Hearing the FOREST

Machine Learning for Biodiversity Monitoring Using Soundscapes

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Business Informatics

by

Dipl.-Ing. Carlos Alberto Vargas Rivera
Registration Number : 11823257 - Sciper : 384891

to the Faculty of Informatics

at the TU Wien

Advisor:	Univ.Prof. Dr. Emanuel Sallinger
Second advisor:	Univ.Prof. Dr. Mathieu Salzmann, (EPFL)
Assistance:	Dr. Carl Remlinger, (EPFL)
	Giacomo Delgado, (ETH Zürich)

Vienna, September 9, 2025

Carlos Alberto Vargas Rivera

Emanuel Sallinger

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Hearing the FOREST

Machine Learning for Biodiversity Monitoring Using Soundscapes

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Wirtschaftsinformatik

eingereicht von

Dipl.-Ing. Carlos Alberto Vargas Rivera
Matrikelnummer : 11823257 - Sciper : 384891

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Prof. Dr. Emanuel Sallinger
Zweitbetreuung: Univ.Prof. Dr. Mathieu Salzmann, (EPFL)
Mitwirkung: Dr. Carl Remlinger, (EPFL)
Giacomo Delgado, (ETH Zürich)

Wien, 9. September 2025

Carlos Alberto Vargas Rivera | Emanuel Sallinger

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Erklärung zur Verfassung der Arbeit

Dipl.-Ing. Carlos Alberto Vargas Rivera

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.
Ich erkläre weiters, dass ich mich generativer KI-Tools lediglich als Hilfsmittel bedient habe und in der vorliegenden Arbeit mein gestalterischer Einfluss überwiegt. Im Anhang „Übersicht verwendeter Hilfsmittel“ habe ich alle generativen KI-Tools gelistet, die verwendet wurden, und angegeben, wo und wie sie verwendet wurden. Für Textpassagen, die ohne substantielle Änderungen übernommen wurden, haben ich jeweils die von mir formulierten Eingaben (Prompts) und die verwendete IT- Anwendung mit ihrem Produktnamen und Versionsnummer/Datum angegeben.

Wien, 9. September 2025

Carlos Alberto Vargas Rivera

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Danksagung

Zunächst möchte ich meiner Familie meinen tiefsten Dank aussprechen, die mich schon in jungen Jahren gelehrt hat, eine respektvolle und harmonische Beziehung zu Pacha-mama und den natürlichen Lebensräumen, die wir auf der Erde teilen, zu pflegen. Ebenso danke ich meinen Freunden, Verwandten und den anonymen Gutachterinnen und Gutachtern, deren Ermutigung und konstruktives Feedback mich während dieses akademischen Weges stets unterstützt haben.

Mein Dank gilt auch Giacomo Delgado, Dr. Carl Remlinger, Dr. Michele Volpi und Univ. Prof. Dr. Mathieu Salzmann für die Zusammenarbeit mit dem Swiss Data Science Center (SDSC), für die Bereitstellung des Rohmaterials, das diese Arbeit erst möglich gemacht hat, sowie für ihre wertvolle Unterstützung und Beratung während des gesamten Forschungsprojekts. Mein herzlicher Dank gilt ferner Univ. Prof. Dr. Emanuel Sallinger für seine Inspiration und dafür, dass er die intellektuelle Neugier gefördert hat, die meine Leidenschaft für die Forschung weiterhin prägt.

Diese Forschung wurde durch finanzielle Mittel und institutionelle Unterstützung ermöglicht. Ich danke für die Bereitstellung der Recheninfrastruktur der HPC-Systeme, die von der EPFL über das von SCITAS betriebene Izar-System sowie über das VSC-5, betrieben vom Austrian Scientific Computing (ASC) an der TU Wien, zur Verfügung gestellt wurden. Weitere finanzielle Unterstützung wurde durch die Kurzfristige Wissenschaftliche Arbeiten (KUWI) [11.022025/1531], das Swiss-European Mobility Programme (SEMP) [384891] und Fundación para el Futuro de Colombia (COLFUTURO) [PCB-18.1144148872] bereitgestellt.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Acknowledgements

First, I would like to express my deepest gratitude to my family, who taught me from an early age to cultivate a respectful and harmonious relationship with Pacha-mama and the natural environments we share on Earth. I am also thankful to my friends, relatives, and the anonymous reviewers whose encouragement and constructive feedback have supported me throughout this academic adventure.

I would like to thank Giacomo Delgado, Dr. Carl Remlinger, Dr. Michele Volpi, and Univ. Prof. Dr. Mathieu Salzmann for the collaboration with the Swiss Data Science Center (SDSC), for sharing the raw material that made this work possible, and for their valuable guidance throughout the research project. I also express my heartfelt gratitude to Univ. Prof. Dr. Emanuel Sallinger for his inspiration and for encouraging the intellectual curiosity that continues to shape my passion for research.

This research was made possible through funding and institutional support. I acknowledge the computational infrastructure of High-Performance Computing systems provided by EPFL via the Izar operated by SCITAS, as well as the VSC-5 operated by Austrian Scientific Computing (ASC) at TU Wien. Additional financial support was provided by the Kurzfristige Wissenschaftliche Arbeiten (KUWI) [11.022025/1531], the Swiss-European Mobility Programme (SEMP) [384891], and Fundación para el Futuro de Colombia (COLFUTURO) [PCB-18.1144148872].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Kurzfassung

Die Agenda 2030 der Vereinten Nationen für nachhaltige Entwicklung hebt die dringende Notwendigkeit hervor, den Verlust der biologischen Vielfalt und die Degradierung von Land zu bekämpfen, die weltweit Ökosysteme und Lebensgrundlagen bedrohen [1]. Diese Arbeit leistet hierzu einen Beitrag, indem sie das Payment for Environmental Services (PSA)-Programm in Costa Rica als Fallstudie für ein groß angelegtes biodiversitätsbezogenes Passive Acoustic Monitoring (PAM) unterstützt.

Wir schlagen das **F**ramew**O**rk for featu**R**e **E**xtraction, vi**S**ualisation, and classifica**T**ion of Soundscapes (FOREST) vor, ein modulares, Python-basiertes Framework, das die Vorverarbeitung, Datensatzkuratierung, Merkmalsextraktion, Visualisierung und prädiktive Klassifikation ökologischer Audioaufnahmen integriert. Zunächst entwickeln wir eine Pipeline, die Klanglandschaften in PyTorch-Tensoren transformiert, einen kuratierten Datensatz der Form $249{,}660 \times 6{,}016$ konsolidiert und fünf (5) statistische Skalare sowie elf (11) Ecological Acoustic Indices (EAIs) extrahiert.

Zweitens entwickeln wir ein Evaluationsframework mit 3,577 experimentellen Durchläufen, um systematisch den Einfluss einzelner Merkmale und ihrer Kombinationen auf die Modellleistung zu analysieren. Die Ergebnisse zeigen, dass eine Teilmenge von fünf (5) EAIs, nämlich NPP, BET, HTP, AEI und HFQ, eine robuste und präzise Klassifikation ermöglicht. Ergänzende Spinnennetz-Visualisierungen zeigen deutliche ökoakustische Profile über die vier (4) untersuchten Ökosystem-Regions (Reference Forest, Pasture, Natural Regeneration, and Plantation) hinweg und stützen die Interpretation spezifischer EAIs als Proxy-Indikatoren für Biodiversität.

Drittens entwerfen und vergleichen wir drei (3) hybride, Deep Learning (DL)-basierte prädiktive Modelle, nämlich ParaNet-CNN-LSTM, SeqNet-CNN-LSTM und SeqNet-LSTM-CNN, mit Basismodellen wie Simple-SVM, Simple-CNN, Simple-LSTM und ResNet1D. Die vergleichende Analyse zeigt, dass ParaNet-CNN-LSTM die konsistenteste und verlässlichste Leistung erzielt, mit Median-Accuracy-Werten über 90% und Maximum-Werten über 97.5% im optimalen Bereich von 10–13 Eingangsmerkmalen.

Das FOREST-Framework fasst diese Beiträge in einer quelloffenen, webbasierten Anwendung zusammen, die unter `www.soundforest.app` zugänglich ist. Auch wenn der Ansatz durch ein unausgeglichenes Datenset, zeitliche Annahmen, fehlende Metadaten und die Einschränkung auf das PSA-Programm begrenzt ist, bietet die Methodik eine

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

solide Grundlage. Diese Arbeit zeigt, dass die Kombination von EAIs mit DL-basierten hybriden Architekturen eine präzise Klassifikation ökologischer Klanglandschaften ermöglicht und einen skalierbaren Ansatz für das zukünftige Biodiversitätsmonitoring in unterschiedlichen Ökosystemen bietet.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Abstract

The 2030 United Nations Agenda for Sustainable Development highlights the urgent need to address biodiversity loss and land degradation, which threaten ecosystems and livelihoods worldwide [1]. This thesis contributes to these efforts by supporting the Payment for Environmental Services (PSA) environmental conservation program in Costa Rica as a case study for large-scale biodiversity Passive Acoustic Monitoring (PAM).

We propose the **F**ramew**O**rk for featu**R**e **E**xtraction, vi**S**ualisation, and classifica**T**ion of Soundscapes (FOREST), a modular Python-based framework that integrates preprocessing, dataset curation, feature extraction, visualisation, and predictive classification of ecological audio recordings. First, we establish a pipeline that transforms soundscapes into PyTorch tensors, consolidating a curated dataset of shape 249,660 × 6,016 and extracting five (5) statistical scalars, and eleven (11) Ecological Acoustic Indices (EAIs).

Second, we develop an evaluation framework comprising 3,577 experimental runs to systematically analyse the impact of individual features and their combinations on model performance. The results show that a subset of five (5) EAIs, namely NPP, BET, HTP, AEI, and HFQ, achieves robust and accurate classification. Complementary spidernet visualisations reveal distinct ecoacoustic profiles across the four (4) studied ecosystem Regions (Reference Forest, Pasture, Natural Regeneration, and Plantation), supporting the interpretation of specific EAIs as proxy-indicators of biodiversity.

Third, we design and benchmark three (3) hybrid Deep Learning (DL)-based predictive models, namely ParaNet-CNN-LSTM, SeqNet-CNN-LSTM, and SeqNet-LSTM-CNN, against baseline models including Simple-SVM, Simple-CNN, Simple-LSTM, and ResNet1D. The comparative analysis shows that ParaNet-CNN-LSTM achieves the most consistent and reliable performance, with median Accuracy values above 90% and Maximum exceeding 97.5% in the optimal range of 10–13 input features set-size.

The FOREST framework consolidates these contributions into an open-source, web-based application accessible at `www.soundforest.app`. While the approach is constrained by dataset imbalance, temporal assumptions, absence of metadata, and its limitation to the PSA program, the methodology provides a rigorous foundation. This thesis demonstrates that combining EAIs with DL-based hybrid architectures enables accurate classification of ecological soundscapes and offers a scalable path for future biodiversity monitoring across diverse ecosystems.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Wien Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Contents

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 1

Introduction

This chapter presents the motivation and context for this research, highlighting the need for effective tools to monitor biodiversity and ecosystem health in response to mitigate global environmental climate-change, and biodiversity loss. It discusses the limitations of traditional survey methods, and introduces Passive Acoustic Monitoring as a scalable, cost-efficient alternative for capturing continuous ecological insights across large geo-spatial and temporal scales.

The aim of the research is to propose a framework for analysing soundscape recordings from diverse ecosystems using Machine Learning techniques. To support this goal, this chapter formulates key Research Questions focused on identifying acoustic patterns across ecological-ecosystems, refining data preprocessing strategies, and designing a predictive Deep Learning model for ecological-soundscapes classification.

In addition, the chapter outlines the methodology used in this research, combining elements of systematic scientific research with software engineering practices to develop practical tools for environmental conservation. Finally, it presents an overview of the tangible contributions and the thesis structure, providing a clear roadmap through the stages of data processing, model design, development, evaluation, and implementation, aimed at supporting biodiversity monitoring.

1.1 Motivation and Problem Statement

The Sustainable Development Goal 15, “Life on Land,” of the 2030 United Nations Agenda for Sustainable Development, is devoted to “Protect, restore and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt and reverse land degradation and halt biodiversity loss” [1]–[3].

While addressing biodiversity loss is an urgent 21st century priority, with land degradation being its largest threat, causing 10% of global emissions and directly impacting 3.2 billion

livelihoods of people [2]–[8]. Combating this issue at a large-scale has proven challenging, given its deep social and ecological complexities [2], [6], [9], [10].

Research shows a direct correlation between ecosystem degradation and wealth inequality. Which leads to overconsumption by the hyper-rich and few non-extractive livelihood options for the extremely poor who are often living in association with nature (i.e. native, or rural-local communities) [6], [11], [12]. Studies suggest redistributing wealth to empower local-rural communities could support reverse ecosystem decline and promote ecological recovery at large-scale [6], [13], [14].

Research shows a direct correlation between ecosystem degradation and wealth inequality, which leads to overconsumption by the hyper-rich and limited non-extractive livelihood options for the extremely poor, often living in close connection with nature (i.e., native or local-rural communities) [6], [11], [12]. Studies suggest that redistributing wealth to empower local-rural communities could help reverse ecosystem decline and promote ecological recovery at large scales [6], [13], [14].

Despite strong theoretical support and tangible environmental outcomes, empirical evidence linking equitable wealth distribution to environmental benefits remains scarce [6], [8], [15], [16]. However, Costa Rica's "Pagos por Servicios Ambientales" [Payment for Environmental Services] (in Spanish: PSA) program offers valuable insights from over 25 years of implementation [6], [17]. PSA compensates local communities for maintaining ecosystem services such as water provision, carbon storage, biodiversity protection, and landscape beauty. As one of the few large-scale Payment for Environmental Services programs with a focus on wealth redistribution, PSA provides a unique case study. Satellite data also indicate that PSA contributes to reducing deforestation and increasing carbon storage [6], [18], [19]. Nevertheless, further research is required to evaluate its broader ecological impacts, given the challenges in collecting comparable landscape-scale ecological data [6], [8], [17].

Tracking biodiversity progress is essential but challenging due to limited resources and the high cost of traditional surveys, which often cover small geographical areas [2], [6], [20]. Passive Acoustic Monitoring (PAM) offers a scalable solution using sound recordings to monitor biodiversity across larger regions. The ecoacoustic approaches are gaining popularity for ecosystem monitoring [21]. Recent studies analyse audio records on a national scale with large datasets to detect ecosystem-level changes [22].

One main challenge in tracking biodiversity is the consistent measurement of ecosystem changes. Ecosystem sounds, also referred to as *Soundscapes*, contain valuable ecological information but are often difficult to interpret directly [6], [23], [24]. To address this issue, researchers have developed Ecological Acoustic Indices (EAIs) to simplify the assessment of biodiversity. However, EAIs are not always reliable and can sometimes yield conflicting results, which limits their usability and credibility [6], [25], [26].

Recent advances in recording technology make it easier to collect large datasets, but the demand for rapid analysis has increased the feasibility of acoustic indices, despite their known limitations. Some researchers argue that focusing these indices on specific

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

components of the soundscape can improve their reliability, although only a few studies analyse sound in detail across both time and frequency [6], [27], [28]. By applying Machine Learning (ML) techniques with Deep Learning (DL)-based models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), and concentrating on the most biologically relevant features of soundscapes, scientists could achieve more accurate monitoring and assessment of biodiversity across diverse ecosystems [6], [12], [20], [26], [29], [30].

1.2 Aim of This Thesis

This research leverages sound recordings from the PSA program to monitor forest ecosystems in Costa Rica. It focuses on detecting and characterising audio signal patterns across four (4) labelled Regions: Reference Forest, Pasture, Natural Regeneration, and Plantation. The primary aim is to model the statistical distribution of the unique acoustic characteristics of soundscapes across these Regions and to train predictive ML-based models capable of classifying them.

To address this goal, a Python-based framework is proposed to process the raw audio samples using state-of-the-art Machine Learning (ML) methods for audio classification, in combination with Music Information Retrieval (MIR) techniques for audio analysis and bioacoustics approaches, including bioacoustic Feature Extraction (FE) and Data Augmentation (DA). The workflow includes filtering noise, segmenting signals by bandwidth, subsampling in time, and enhancing key properties of the raw recordings to reveal frequency- and time-based patterns, enabling the training of predictive models that classify the regional origin of new audio samples.

Furthermore, Deep Learning (DL) architectures, including CNN and RNN, are applied to the pre-processed soundscapes to assess their similarity to a baseline region and rank them accordingly. The work addresses challenges such as handling high-dimensional data on High-Performance Computing (HPC) systems, optimising Python-based processing pipelines, and exploring alternative model architectures to improve classification accuracy and robustness.

The main contribution of this work is the proposal of the Python-based **F**ramew**O**rk for featu**R**e **E**xtraction, vi**S**ualisation, and classifica**T**ion of Soundscapes (FOREST).

1.3 Methodology

This section presents the multi-method research approach adopted in this thesis to address the Research Questions about soundscape analysis and ecosystem classification. It combines established scientific methods such Design Science Research Method, Systematic Literature Review, and Cross-Industry Standard Process for Data Mining with SCRUM Agile for Software Development. This integrated methodology ensures a

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

rigorous, systematic exploration of the Research Questions while enabling the iterative development of practical software artifacts to support biodiversity monitoring efforts.

1.3.1 Research Questions (RQs)

There are three Research Questions (RQs) that aim to guide the thesis by identifying the key characteristics of audio time series across different ecosystem regions, determining effective data preprocessing and augmentation techniques to enhance model training, and designing an appropriate deep learning model architecture for accurately classifying soundscapes according to the ecosystem Regions. These questions structure the research focus and methodology throughout the work.

RQ1 What are the appropriate data pre-processing and Data Augmentation techniques that can be applied to raw audio recordings (*soundscapes*) to perform model training, and impact predictive performance?

RQ2 What are the most relevant characteristics of the soundscapes among the ecosystem Regions (Reference Forest, Pasture, Natural Regeneration, and Plantation)?

RQ3 What is an appropriate Deep Learning model architecture to learn the statistical patterns of soundscapes and accurately classify audio signals among the different environmental ecosystem Regions?

1.3.2 Research Methodology

To address the research gap identified in this master thesis, we adopt a multi-method research approach that combines both established scientific methods and practical development frameworks. This approach ensures that the Research Questions are systematically explored while allowing the creation of working software artifacts throughout the research project lifecycle.

Three (3) main research methods structure the scientific study, complemented by one (1) secondary method that supports the software development process:

- Design Science Research Method (DSRM)
- Systematic Literature Review (SLR)
- Cross-Industry Standard Process for Data Mining (CRISP-DM)
- SCRUM Agile for Software Development (SCRUM) (secondary)

Together, these methods provide a complementary foundation for the thesis: DSRM structures the design and evaluation process, SLR builds the theoretical background, and CRISP-DM supports systematic data handling. Finally, SCRUM supports the iterative development of software artifacts, ensuring that research outputs remain aligned with practical requirements.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Design Science Research Method (DSRM)

The DSRM provides the overarching framework for guiding the design, implementation, evaluation, and communication of technological artifacts in this thesis [31]–[33]. It offers a structured problem-solving approach that emphasises both theoretical contribution and practical relevance, and it is organised into six (6) phases:

1. Problem Identification - Conduct interviews and contextual analysis of conservation needs and bioacoustic workflows.
2. Define Objectives - Establish technical and practical success criteria.
3. Design & Development - Iteratively create the proposed framework, including soundscape processing and predictive classification components.
4. Demonstration & Implementation - Develop use-case prototypes and proof-of-concepts with real-world datasets.
5. Evaluation - Perform systematic model benchmarking and collect expert results.
6. Communication - Disseminate results through this thesis and deploy the application at `www.soundforest.app`.

The artifacts generated during these phases (software, models, datasets, documentation) are listed in Appendix A.

Systematic Literature Review (SLR)

The SLR collects, evaluates, and synthesises relevant scientific literature aligned with our research questions. This method follows a rigorous, transparent, and replicable process adapted for the Information Systems and Data Science domains [34]–[38].

In the context of this thesis, the SLR provides a solid foundation for identifying novel techniques, methodological patterns, and insights relevant to domains such as bioacoustics, biodiversity conservation, and environmental audio processing. The detailed protocol, databases, and the strategy to perform the SLR appear in Appendix B.

Topics Covered per Research Question:

RQ1 Data Preprocessing, Data Transformation, Data Augmentation (DA), Music Information Retrieval (MIR), Ecological Acoustic Indices (EAIs)

RQ2 Data Exploratory Analysis (DEA), Cluster Analysis, Feature Extraction (FE), Factor Analysis (FA), Data Visualisation

RQ3 Machine Learning (ML) frameworks Classification, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), Classification Metrics, Model Evaluation, Federal Learning, Datasets

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Databases Accessed: *Österreichische Dissertationsdatenbank, IEEE/IET Electronic Library (IEL), ACM Digital Library, arXiv, ScienceDirect, Springer Link, Scopus*

Cross-Industry Standard Process for Data Mining (CRISP-DM)

Since the DSRM provides the research-framework guidance, the CRISP-DM structures the data-related processes of this thesis. It acts as the main methodological foundation for DEA, data preprocessing, and predictive model design, tailored to the requirements of bioacoustics and environmental conservation [39], [40]. The CRISP-DM defines the following six (6) phases:

1. Business Understanding – Align the goals of the project with domain experts in biodiversity monitoring and environmental conservation.
2. Data Understanding – Apply DEA to soundscape datasets to assess raw data characteristics, statistical distributions, and quality, and to identify initial patterns.
3. Data Preparation – Clean, transform, augment, and extract features and EAIs from raw audio data.
4. Modelling – Design, develop, and train ML models suited for time-series and acoustic classification tasks.
5. Evaluation – Benchmark the FOREST proposed models performance against SOTA models.
6. Deployment – Integrate and deploy the resulting artifacts into the open-source application available at `www.soundforest.app`.

The list of artifacts generated through the CRISP-DM process appears in Appendix A.

SCRUM Agile for Software Development (SCRUM)

To manage the practical working software development of the FOREST, the SCRUM framework is utilised. The fundamental Agile principles provide the necessary flexibility to adapt to change under time pressure, respond to stakeholder feedback, and iteratively improve deliverables [41], [42].

The software development lifecycle follows these core phases:

- Planning and Analysis – Define a product backlog based on stakeholder interviews, and Systematic Literature Review.
- Design and Development – Execute two-week sprints with tasks prioritised and rescheduled according to evolving requirements and iterative feedback.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

- Implementation – Deliver in three stages: proof of concept, functional prototype, and final deployment.

This lightweight yet well-adopted methodology allows the project to evolve continuously toward a robust, production-ready working solution. Incremental outputs are reviewed and adjusted regularly, ensuring alignment with both technical goals and the environmental conservation use case [41], [42].

In summary, the integration of DSRM, SLR, CRISP-DM, and SCRUM provides a comprehensive methodology for navigating scientific investigation and real-world software development, ensuring methodological rigour while delivering impactful technological artifacts that support environmental conservation, and address biodiversity loss.

1.4 Main Contribution

This research contributes to the development of scalable methods and tools for soundscape analysis in the context of biodiversity monitoring and environmental conservation based on Passive Acoustic Monitoring (PAM). The work addresses the design of a methodological framework, the development of a preprocessing and data-transformation pipeline, the generation and visualisation of curated datasets, the proposal of DL-based models for predictive audio classification, the introduction of a robust set of Ecological Acoustic Indices (EAIs) that optimise the soundscape classification task, and the release of the Python-based open-source FOREST. The specific contributions are the following.

Systematic Literature Review (SLR). A structured SLR of existing scientific literature is conducted to identify relevant methods, methodological patterns, and research gaps in soundscape analysis, bioacoustics, and ML-based methods for audio classification. The SLR provides the theoretical foundation for the design of the proposed framework.

Data processing pipeline. A novel Python-based pipeline is developed to preprocess raw soundscapes through noise filtering, temporal and spectral segmentation, and data transformation. The pipeline integrates Feature Extraction (FE), Data Augmentation (DA), and the visualisation of EAIs, generating interpretable outputs that support both environmental experts, and data scientists.

Curated dataset. A dataset of shape $249{,}660 \times 6{,}016$ is generated from raw soundscapes collected across four (4) labelled Regions. The dataset incorporates systematic cleaning, augmentation, and Feature Extraction (FE), as well as the integration of Ecological Acoustic Indices (EAIs), providing a reproducible foundation for large-scale biodiversity monitoring experiments.

Data visualisation. A visualisation strategy is proposed to represent the EAIs extracted from each soundscape, emphasising unique acoustic characteristics that distinguish

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

each specific environmental regions. This facilitates the interpretation of audio features, supports exploratory analysis, and provides insights into ecosystem-specific acoustic patterns.

Predictive models. A family of three Deep Learning (DL)-based hybrid architectures, namely Parallel Network CNN - LSTM (ParaNet-CNN-LSTM), Sequential Network CNN - LSTM (SeqNet-CNN-LSTM), and Sequential Network LSTM - CNN (SeqNet-LSTM-CNN), is designed and implemented for the soundscape classification task. These models build upon Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) components, are trained and validated using the curated dataset, and are benchmarked against State of the art (SOTA) approaches in audio classification.

Evaluation framework. A systematic evaluation methodology is developed to assess predictive performance. It examines the effect of the number of EAIs, compares their impact across different architectures, and quantifies the individual contribution of each EAI to model accuracy. In total, 3,577 experiments are executed on High-Performance Computing (HPC) systems to ensure scalability and robustness.

Model selection. Based on the evaluation framework, the most suitable DL-based architecture is identified. The optimised model is proposed as the best-performing solution for large-scale soundscape classification in biodiversity monitoring.

Robust feature set selection. Based on the evaluation framework, the individual contributions of each EAI are assessed to propose a robust subset of EAIs that optimise predictive performance. This supports more efficient model training and enhances the interpretability of acoustic features for further ecological applications.

Software application. The **F**ramew**O**rk for featu**R**e **E**xtraction, vi**S**ualisation, and classifica**T**ion of Soundscapes (FOREST) encapsulates all contributions into a modular and extensible Python-based system. It integrates preprocessing, dataset curation, feature extraction, visualisation, predictive modelling, and evaluation into a unified workflow. The system is deployed as an open-source, web-based application accessible at `www.soundforest.app`.

In summary, these contributions establish a comprehensive framework for soundscape analysis, combining methodological innovation, data curation, predictive modelling, data mining, and software engineering. By integrating preprocessing, feature engineering, visualisation, evaluation, and deployment into a unified workflow, this research advances the State of the art in bioacoustic monitoring and provides interpretable tools to support large-scale biodiversity conservation efforts.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

1.5 Structure of This Thesis

This thesis is structured to guide the reader through the full research journey from the theoretical foundation to the final conclusion, with a focus on designing, evaluating, and proposing a novel predictive model for soundscape classification.

In Chapter 2, the data preprocessing pipeline is described in detail. This includes techniques for sampling, cleaning, transforming, and normalising the data, as well as Feature Extraction (FE) and Data Augmentation (DA) strategies tailored to the specific challenges of the bioacoustic domain. The chapter concludes with a description of the curated dataset that results from this process, which serves as the input for model training.

In Chapter 3, the design of the predictive models is presented. The chapter outlines the model architectures, the training strategy, and introduces three (3) Deep Learning (DL)-based models for the soundscape classification task. It also explains how the curated dataset is used during both the training and evaluation stages. Finally, the chapter describes the metrics and procedures applied to select the most appropriate model among the proposed architectures, benchmark them against SOTA models, and identify an optimised model for the classification task.

In Chapter 4, the performance of the proposed models is evaluated and compared with other SOTA approaches in the domain of audio classification. The experimental configuration is first presented, followed by a step-by-step analysis of the results: assessing the impact of the number of Ecological Acoustic Indices (EAIs) across different model architectures, identifying the most appropriate architecture for the soundscape classification task, and evaluating the individual contribution of each EAI to model performance. Finally, the chapter proposes a particular model architecture and a set of EAIs that optimise the performance of the soundscape classification task.

Finally, Chapter 5 brings the research to a close. The last chapter provides a discussion of the key findings, acknowledges the limitations of the current work, and delivers a comprehensive conclusion. The thesis ends with suggestions for future research directions that build on the presented work.

References appear in the appendices of the document, supporting the theoretical background, methodology, project timeline, specific delivered artifacts, and discussions throughout the thesis.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 2

Soundscapes Preprocessing

Chapter 2 presents the preprocessing pipeline for soundscapes, a foundational step in providing effective bioacoustic analysis. Section 2.1 details the audio preprocessing phase, starting with the loading of raw soundscape records. Then, the section also introduces the subsampling (framing) strategy, in which audio recordings are segmented into overlapping, and non-overlapping frames to enable efficient and scalable data handling for noise reduction, and spectral analysis.

Section 2.2 enumerates and describes the computation of the Ecological Acoustic Indice (EAI) that serve as proxies for biodiversity and ecosystem activity. Section 2.3 outlines the construction of the curated dataset resulting from the preprocessing steps, highlighting its structure and readiness for downstream modelling tasks.

Finally, Section 2.3 presents the complete preprocessing and data transformation workflow, detailing the sequential steps taken to convert raw audio data into a structured format suitable for further computational tasks. This includes the description of the curated dataset and a graphical overview highlighting selected EAIs obtained from the processed soundscapes.

2.1 Audio Preprocessing

This section introduces the initial steps for handling raw soundscape recordings. It covers loading high-quality audio files with preserved sample rates, and subsampling the audio into overlapping, and non-overlapping frames. Furthermore, it exposes the preprocessing set of tasks, and data transformation strategies to ensure that the raw soundscapes are transformed into a suitable format for efficient processing and model training.

2.1.1 Loading Soundscapes

This study used a dataset that has been collected under the Payment for Environmental Services (PSA). Each soundscape is an audio record with sounds of forests' ecosystems utilising the Passive Acoustic Monitoring (PAM) technique. Each record has been labelled according to one of the four regional land cover types: Reference Forest, Pasture, Natural Regeneration, and Plantation.

The raw records have a duration of 30 minutes (1800 seconds) with a total size of 14.4MB per soundscape. Each soundscape corresponds to one of the four labelled Regions (Reference Forest, Pasture, Natural Regeneration, and Plantation). While using the Librosa python package to load the soundscapes with the default settings, we preserve the original sample rate of 48kHz (with the setting: sr = None), for all the other settings we assume the default values (see Librosa.load() function documentation). These higher frequencies are crucial as they encompass the vocalizations of certain insects, the upper harmonics of some bird calls, and the echolocation sounds produced by bats at higher frequencies.

In our recordings, we employed a bit depth of 48 bits, which strikes a balance between audio quality and data size in mp3 audio format as it was given by the dataset. It is important to mention that one of the reasons to use a digitalised file format for the recordings in this format, is due to the capabilities for the Librosa library to manipulate the records and generate artifacts such as the spectrograms. Furthermore, we emphasise the implementation of additional libraries such Multiresolution Analysis of Acoustic Diversity (MAAD) [43], and PyTorch [44] to accurately, and efficiently compute a set of Ecological Acoustic Indice (EAI), and perform Data Augmentation (DA) tasks, which are critical for our analysis.

2.1.2 Framing Without Overlapping

For the framing without overlapping approach we considered soundscapes digitalised at 48000 samples per second and subdivided into non-overlapping frames, each containing 288000 samples, we end up with 300 frames in total, each of these frames essentially holds a time duration of 6 seconds. These frames are used for further processing [27].

2.1.3 Wave Envelope

The wave envelope is calculated from the audio signal waveform, we take a methodological approach by gathering the maximum absolute amplitude present within each defined frame of the recording (without overlapping) [27]. The total number of values representing the wave envelope for a typical 6 seconds segment of audio is calculated by taking the maximum absolute of a subsample of 20 milliseconds among the 6 seconds (6000 milliseconds) frames. This strategy directly correlates with the number of complete frames processed within that time span, and reduces the dimensionality of the 6 seconds frames of 288000 samples each to a 5760 samples per 6 seconds for the whole 30 minutes soundscapes (300 frames of 6 seconds).

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

To convert the absolute amplitude values obtained from the wave envelope into a more useful format for analysis, we express these values in decibels (dB) using the following formula: dB = 20 × log10(A), where A represents the amplitude value. To mitigate any potential errors stemming from zero signal values. We establish a minimum dB threshold value set at -90 dB. While this minimum value is somewhat high, it is regarded as suitable for the types of acoustic recordings [27]. However, it is important to highlight that this minimum dB threshold should be adaptable based on various factors such as the quality of the microphones used, the bitrate of the recordings, and other considerations impacting the overall audio quality [27]. These frames are used for further processing.

2.1.4 Noise Reduction

In order to reduce the noise of the soundscape signals. We implemented two different noise reduction approaches: stationary noise reduction and non-stationary noise reduction. The first strategy applies noise reduction assuming that the background noise remains constant over time (nr.reduce_noise library) [45], [46] with the stationary=True flag and a noise threshold parameter to 1.9 standard deviations that determines how aggressively noise is filtered out [27]. The second strategy applies a different noise reduction method that assumes the noise varies over time, using a slightly different threshold to 1.6 standard deviations and setting stationary=False. We only reduce the noise of the non-overlapping vectors, and each frame of 6 seconds is passed through both denoising methods. Once processed, the results are stored as NumPy arrays. These denoise frames are used for further processing, and hold the same dimensionality of the non-overlapping frames. Thus, 288000 samples per 6 seconds among 300 frames.

Noise Reduction via Spectral Gating

The noise reduction algorithm is based on spectral gating, where the Short-Time Fourier Transform (STFT) of the input signal is processed to suppress noise-dominated time-frequency bins. The denoised signal $\hat{Y}(f,t)$ is computed as:

$$\hat{Y}(f,t) = \begin{cases} Y(f,t) \cdot (1 - \alpha \cdot M(f,t)), & \text{if } M(f,t) < \theta. \\ Y(f,t), & \text{otherwise.} \end{cases} \tag{2.1}$$

Noise Reduction Formula and Cases [45], [46]

Where:

- $Y(f,t)$ is the STFT of the input signal $y(t)$ at frequency bin f and time frame t.
- $\hat{Y}(f,t)$ is the denoised STFT.
- $M(f,t)$ is the mask value indicating the likelihood of noise at each time-frequency bin.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

- $\alpha \in [0, 1]$ is the noise reduction proportion (parameter `prop_decrease`).
- θ is the threshold that determines whether a bin is noise-dominated.

The noise mask $M(f, t)$ is computed using a sigmoid-based thresholding of the spectral difference between the signal and estimated noise:

$$M(f,t) = \sigma\left(\frac{|Y(f,t)| - \mu_N(f)}{\sigma_N(f)} \cdot \gamma\right), \tag{2.2}$$

$$\mu_N(f) = \mathbb{E}[|N(f,t)|], \tag{2.3}$$

$$\sigma_N(f) = \text{StdDev}[|N(f,t)|]. \tag{2.4}$$

Noise Calculation Function [45], [46]

Where:

- $N(f, t)$ is the STFT of the noise signal $y_{\text{noise}}(t)$, when provided.
- $\mu_N(f)$ and $\sigma_N(f)$ are the mean and standard deviation of the noise spectrum at frequency f.
- γ is the sigmoid slope (parameter `sigmoid_slope_nonstationary`).
- $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function used for smooth mask transitions.

The final denoised signal is reconstructed via inverse STFT of $\hat{Y}(f, t)$.

2.1.5 Spectrograms

Yoh et al. have proposed a method to leverage soundscapes, and similar cyclic data using Fast Fourier Transform (FFT) and the Spectrograms properties [47] and find patterns in the composition and decomposition of the signals along the time. Therefore, this is also a main step for our research. To generate the spectrograms we used the overlapping frames while applying the well-known Hanning windows strategy (438 frames per 6 seconds) [48]–[52]. Then, a Fast Fourier Transform switches from the time domain to the frequency domain, setting the limits of the lower band to 0 and the upper band to 10000 using the butter bandstop filter function from the library maad (butter_bandstop_filter) to filter the samples per desired frequency band [27]. Depending on the requirements of the subsequent analyses to be performed, the spectrograms are computed using the sound.spectogram function from the package maad with setting mode to amplitude.

The output of this spectrogram generation strategy is represented by a tuple with the following vectors: Sxx, tn, fn, and ext. Sxx represents the spectrogram of an audio

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

signal, which is a visual and numerical representation of how the frequency content of the sound changes over time as a 2-Dimensional array that holds the spectrogram data. Each column represents a time segment, and each row represents a frequency bin. The values in this matrix indicate the intensity of the amplitude (the higher the more intense) of the frequencies at different time points. The tn is the array representing the time values corresponding to the spectrogram's columns, indicating when each segment was analysed. The fn is an array containing the frequency values, representing the different frequency components analysed in the signal. Finally the ext array contains additional metadata, and it is not further used.

Overall, this output allows for the analysis of how different frequencies are distributed across time in an audio signal and are used for further processing of the Ecological Acoustic Indice (EAI).

2.1.6 Framing with Overlapping

The framing with overlapping is a well-know strategy to subsample signals over the time or the frequency dimensions. Müller et al. implemented a three-second audio vectors strategy using a Short-Time Fourier Transform (STFT) with a window size of 0.1-second and a hop length of 0.025-second to create spectrograms. Their frequency range was set between 50 Hz and 8 kHz, divided into 224 frequency bins. The resulting amplitudes were log-scaled, and the frequency axis was transformed using the mel-scale [26]. Furthermore, Tom Bradfer-Lawrence et al. used raw recordings of 10-minutes for 3-years records, and subsampled them into 1-minute vectors [53].

2.1.7 Hanning Window

The Hann window has its origins in the work of Julius von Hann (1839–1921), a pioneering Austrian meteorologist regarded as a founder of modern meteorology. Hann studied mathematics, chemistry, and physics at the University of Vienna, where he later became a professor of meteorology, and cosmic physics. He also served as director of the Central Institute for Meteorology and Earth Magnetism in Vienna. Among his contributions, Hann developed a weighted smoothing technique for combining meteorological data from neighbouring regions, using weights of [1/4, 1/2, 1/4], a method now known as Hann smoothing [48]. This technique was later adapted by Blackman and Tukey in 1961 to define the Hanning window function for spectral analysis [48], [49]. Nowadays, this is a widely used window function for signal processing in different domains [50]–[52].

2.1.8 Framing with Overlapping and Hanning Window

For this research, we decided to implement a much shorter subsample strategy compared with the used by Müller et at. and Bradfer-Lawrence et at. [26], [53] due that we aim to support records of less than one (1) minute to use as input external data for the preprocessing pipelines. However, the process is identical as many strategies already implemented in the literature [13], [16], [26], [27], [50]–[52], [54], [55]. First, we determine

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

the size of each frame size calculated as the sample rate per the total seconds of the frame resulting in frames of 288000 samples. Since by definition the Hanning windows is used for overlapping frames. Then, the hop size of the frames are calculated with an overlapping of 1.9 seconds. These values are global constants that can be specified at the running time, and will propagate along the whole framework. Nonetheless, are explicitly as explained in this section. The total number of frames is then 438 instead of the non-overlapping strategy that outputs 300.

The frames are then calculated on a function that takes the whole soundscape audio signal and divides it into frames. If the Hanning window is enabled (see np.hanning), each frame is multiplied by a windowing function that gradually tapers the edges, reducing artifacts that can occur when analysing the frequency content of the audio. Without this window, the transitions between frames might introduce unwanted distortions. The function extracts these frames by shifting through the signal in steps, ensuring that the frames overlap, which helps preserve information that might otherwise be lost if each segment was completely separate. Finally, the function returns two copies of the framed audio signal, producing two lists of overlapping audio frames: one that applies the Hanning function, and the other list only with the overlapping frames without any applied function. These frames are used for further processing.

2.1.9 FFT with Overlapping Hanning Window Frames

Following the strategy implemented by Müller et al. [26]. One of the most important reasons for using the Hanning window framing strategy on our proposed framework is the adequate performance, and efficacy at splitting the signal using the FFT to each frame. We implemented the Numpy function np.fft.fftfreq that generates an array of frequencies corresponding to the Fourier-transformed signal which is crucial for converting the FFT bin indices into actual frequency values. The output of np.fft.fftfreq consists of frequency values ranging from 0 up to the Nyquist frequency (sr/2), followed by negative frequencies that wrap around due to the symmetric properties of the FFT.

The FFT itself is applied to the segmented signal using np.fft.fft, producing a complex-valued spectrum representing both amplitude and phase information for each frequency component. The real and imaginary parts of the spectrum encode different aspects of the signal's frequency content. Once the FFT is computed, our split band function isolates specific frequency bands by creating masks based on the output of np.fft.fftfreq. The mask array filters the FFT output, selecting only the components within a given frequency range (from the given lower and upper frequency bands with a step bandwidth) similar to the analysis made by Villanueva Rivera et al. [56], [57]. The selected components are then transformed back into the time domain using the Inverse Fast Fourier Transform (IFFT) (np.fft.ifft), yielding band-limited versions of the original signal.

The final output consists of a list of 438 np.ndarrays having n=10 for each of the split bands from the lower band limit of 0 up to the upper band limit of 10000 with bandwidth 1000. Each containing a band-limited reconstruction of the original signal in the time

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

domain. These reconstructed signals retain only the amplitude of the signal (after the Hanning transformation has been applied) within the specified frequency bands, allowing further analysis of different acoustic components.

The resulting vectors contain real-valued time-domain signals after extracting relevant frequency components and discarding the imaginary part with a dimension of 6000 for a 6 seconds record. Since this function is computationally demanding, each of the 438 audio frames is processed in parallel, the implementation ensures efficient computation across multiple cores, significantly reducing the total processing time. These vectors are used for further processing of the Bioacoustic Indices.

2.2 Ecological Acoustic Indices

The Ecological Acoustic Indice (EAI) are widely used to study soundscapes due to their simplicity and assumed connection to biological activity [58]. They are particularly useful in large-scale research that involves many recording devices, which has become increasingly common with the rise of affordable and accessible acoustic sensors. However, differences in how these sensors capture sound can introduce a new source of variability that is often overlooked. This variation may influence results in studies that assume sound differences come only from spatial or temporal factors, highlighting the importance of understanding how sensor characteristics can shape acoustic data [27], [54].

Recent researches have used the EAI as proxies for the biodiversity analysis [13], [23], [26], [27], [54], [55], [59]. Within our proposed framework, we systematically compute various bioacoustic indices that serve as analytical tools for examining acoustic data. These indices provide insights into the soundscape's properties and the dynamics of the acoustic events captured in our recordings as a whole. They summarise with a scalar the specific characteristic of the given array (frame), and are computed either among the whole spectrum of the signal, and the specific frequency band. We aim to expose a brief interpretation of each of the EAI to give more context along the research as an interpretability outcome [54], [55].

After the initial preprocessing phase of the soundscapes, the Multiresolution Analysis of Acoustic Diversity (MAAD) library is implemented for the calculation of most EAIs. MAAD is an open-source Python package developed by Colombian and French Environmental Research Institutions and funded by the Colombian Ministry of Environment and Sustainable Development. It provides a marvellous functionality for acoustic Feature Extraction and serves as a reliable open-source solution for soundscape analysis [43].

Acoustic Complexity Index (ACI)

Depraetere et al. proposed this index to quantify the variation of sound intensity over time [60]. It detects fluctuations in amplitude, which are typical of natural environments where sounds (e.g., bird songs, insect calls) vary in intensity [61], [62]. High ACI values suggest a diverse and dynamic soundscape, while low values indicate constant,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

monotonous noise (e.g., traffic, farming noise). While monitoring ecosystem health, a decreasing ACI over time may suggest habitat degradation. At detecting human-made noise, low ACI values in urban areas indicate dominance of constant sounds like motors or machinery.

ACI calculation operates in the time-frequency domain, and its analytical basis lies in the spectrogram matrix S. This matrix is composed of Fourier coefficients, which encode the signal intensity across time and frequency bins. The computation involves the following steps:

1. Compute the STFT of the input signal to obtain the spectrogram matrix S.
2. Calculate the absolute difference of intensity values in adjacent time bins:

$$|S_{n,f} - S_{n+1,f}|.$$

   This absolute difference between the adjacent bins can be interpreted as the temporal derivative of the spectrogram.

3. Normalize each difference by the total intensity in that frequency band:

$$\frac{|S_{n,f} - S_{n+1,f}|}{\sum_{n=0}^{N_t-1} S_{n,f}}.$$

   This step ensures that the index is relative and reduces the influence of signal amplitude (e.g., due to distance from the microphone).

4. Sum the normalised differences over all frequency bins and time steps to obtain the final ACI score:

$$\text{ACI} = \sum_{f=0}^{N_f-1} \left(\sum_{n=0}^{N_t-2} \frac{|S_{n,f} - S_{n+1,f}|}{\sum_{n=0}^{N_t-1} S_{n,f}} \right). \tag{2.5}$$

Equation for Acoustic Complexity Index (ACI) [61]

Where:

- $S_{n,f}$ is the spectrogram value at time index n and frequency index f.
- N_t is the number of time frames.
- N_f is the number of frequency bins.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Alternative Acoustic Complexity Index (ACA)

The Alternative Acoustic Complexity Index (ACA) [60] is a variation of the Acoustic Complexity Index (ACI) [61], calculated with a different equation based on the original mathematical principal. The goal is to use the ACA as a comparative metric alongside ACI and validate results.

This alternative formulation first computes the Acoustic Complexity Index across the frequency dimension, followed by the time dimension, using efficient vectorised operations with NumPy as follows:

- Normalize the spectrogram matrix S by its maximum value:

$$S' = \frac{S}{\max(S)}.$$

- Calculate the absolute difference of intensity values in adjacent frequency bins:

$$|S'_{f+1,t} - S'_{f,t}|.$$

- Normalize each difference by the sum of intensities of the adjacent bins:

$$\frac{|S'_{f+1,t} - S'_{f,t}|}{S'_{f+1,t} + S'_{f,t}}.$$

- Sum the normalised differences over all frequency bins and time steps to obtain the final frequency-time Alternative Acoustic Complexity Index (ACA_{ft}):

$$\mathrm{ACA}_{ft} = \sum_{t=0}^{N_t-1} \sum_{f=0}^{N_f-2} \frac{|S'_{f+1,t} - S'_{f,t}|}{S'_{f+1,t} + S'_{f,t}}. \tag{2.6}$$

Equation for frequency-time Alternative Acoustic Complexity Index (ACA_{ft})

Where:

- $S'_{f,t}$ is the normalised spectrogram value at frequency bin f and time frame t.
- N_f is the total number of frequency bins.
- N_t is the total number of time frames.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Acoustic Diversity Index (ADI)

The Acoustic Diversity Index (ADI) has been developed by Villanueva-Rivera et al. [57] and further research by Pieretti et al. [61] ADI evaluates how sound energy is distributed across different frequency bands [56], [57], [59], [61], [63]. The higher the ADI suggests many active frequency bands, indicating a rich and diverse soundscape. The lower the ADI suggests sound energy is concentrated in a few bands, often due to dominant noise sources (e.g., industrial noise or a single bird species vocalizing). It is one of the most important metrics to Identify diverse ecosystems with many species vocalizing at different frequencies. At detecting habitat degradation, where a loss of species results in fewer active frequency bands.

The Acoustic Diversity Index computation involves the following steps:

1. Normalize the spectrogram S by its maximum value:

$$S_{\text{norm}} = \frac{S}{\max(S)}.$$

2. Convert the background noise threshold bn from decibels (dBFS) to a linear scale:

$$bn = 10^{\frac{bn}{20}}.$$

3. Subtract the threshold from the spectrogram and binarize the result:

$$S_{\text{clean}} = \max(0, S_{\text{norm}} - bn).$$

$$S_{\text{binary}}(i,j) = \begin{cases} 1, & \text{if } S_{\text{clean}}(i,j) > 0. \\ 0, & \text{otherwise.} \end{cases}$$

4. Divide the frequency range up to $F_{\max}$ into n bands of width w_{band}:

$$n = \left\lfloor \frac{F_{\max}}{w_{\text{band}}} \right\rfloor.$$

5. For each band i, compute the relative activity p_i as the proportion of active (nonzero) bins:

$$p_i = \frac{\sum S_{\text{binary}}(f,t)}{N} \quad \text{, for frequencies } f \in \text{band } i.$$

   Where N is the total number of bins in the band.

6. Add a small epsilon to avoid logarithm of zero, and compute Shannon entropy:

$$ADI = -\sum_{i=1}^{n} p_i \cdot \log(p_i). \tag{2.7}$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Equation for Acoustic Diversity Index (ADI) [56], [57]

Where the following are the parameters utilised per frame:

- S: Spectrogram of the audio signal.
- F_{max}: Maximum frequency considered (default: 10,000 Hz).
- w_{band}: Width of each frequency band (default: 1,000 Hz).
- bn: Background noise threshold in dBFS (default: -50 dBFS).

Bioacoustic Index (BET)

The Bioacoustic Index focuses on sound energy levels in the 2–8 kHz range, where most bird and insect calls occur. The BET has been developed by Boelman et al. [63] and further researched by Pieretti et al, and Villanueva-Rivera et al. [56], [57], [59], [61]. It provides a single scalar that represents biological activity. Higher values indicate high bioacoustic activity, while lower values suggest reduced biodiversity or silence. When monitoring bird and insect populations in different habitats we assume that all the soundscapes are recorded at the same daytime due to this, and other indices are sensible comparing day vs. night soundscapes (e.g., nocturnal vs. diurnal species). The BET can be calculated as follow:

$$\text{BET} = \sum_{f=f_1}^{f_2} s(f). \tag{2.8}$$

Equation for Bioacoustic Index (BET) [43], [63]

Where:

- $s(f)$ is the mean spectral amplitude in decibels (dB) at frequency f.
- The frequency bounds used for this research are:

$$f_1 = 0 \text{ Hz}, \quad f_2 = 10000 \text{ Hz}.$$

Temporal Median (MMM)

The Temporal Median calculates the median amplitude of the audio signal over time, it was been utilised by Depraetere et al. in their researches [60]. Unlike the mean (which can be affected by extreme values), the median is more robust against outliers. It helps in understanding background noise levels and overall loudness stability [27], [60]. When comparing soundscapes: A low median suggests a quiet environment, while a high median

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

suggests continuous sound. Also it is appropriate for Identifying sudden loud events: A big difference between the median and max amplitude suggests short, intense sounds (e.g., rosters, lions). The MMM can be calculated as:

$$MMM = \text{median}(a(n)) \cdot 2^{(1-d)}. \tag{2.9}$$

Equation for Temporal Median (MMM) [43], [60]

Where $a(n) \in \mathbb{R}^{N_t}$ is the amplitude envelope of the signal and d is a dimensionless damping factor (commonly related to the dynamics or noise level of the signal). For this research we used the Hilbert transform [64] to extract $a(n)$, which captures the instantaneous amplitude of the signal.

Number of Peaks (NPP)

The Number of Peaks Index counts the number of distinct peaks in the spectrogram [43]. The Peaks represent individual sound events, such as: Bird calls, insect chirps, or amphibian croaks. Also human-made noises like honks, alarms, or mechanical beeps. Gasc et al. have analysed the NPP as a biodiversity indicator and found that the level of animal activity determined by ear and the number of NPP measured on the mean spectrum before ambient noise reduction highlighted an increase of NPP with the level of animal activity [65]. Therefore, the higher the number of peaks suggests a richer and more dynamic soundscape, while the lower count may indicate background noise or silence. When estimating animal activity, the more the peaks suggest many vocalizations. At detecting mechanical sounds, repetitive peak patterns can indicate farm noise, alarms, or construction activity.

The NPP Index quantifies acoustic richness by counting prominent peaks in the average spectrum of a recording. The following algorithm outlines the computation steps [66].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Algorithm 2.1: Computation of the Number of Peaks (NPP) [43], [65], [66]

Input: Audio signal $x(t)$, sample rate f_s
Output: Number of Peaks (NPP)

```
begin
    Compute the mean spectrum S(f) using a method such as FFT-based
    Normalize the spectrum to the range [0, 1]: S_norm(f) = (S(f) − min(S)) / (max(S) − min(S))
    Apply a peak detection algorithm on S_norm(f) to find local maxima
    foreach detected peak at frequency f_i do
        if Slope of peak > 0.01 amplitude units and distance to adjacent peaks
          ≥ 200 Hz then
            Keep the peak
        end
        else
            Discard the peak
        end
    end
    NPP ← total number of valid peaks
end
```

Frequency Entropy (HFQ)

The Frequency Entropy has been proposed, and utilised by Sueur et al [67]. It measures the randomness or uniformity of a system and describes how evenly energy (amplitude) is spread across frequencies [67]. The lower the entropy the fewer frequencies dominate among the others (e.g., a single species tones, or farming) [27]. A high entropy means energy is spread across many frequencies (e.g., rain, wind, bird chorus). It can help to differentiate between natural and artificial sounds at identifying monotone noises (low entropy) vs. diverse soundscapes (high entropy).

The Frequency Entropy (H_f) also known as Spectral Entropy Index H_f quantifies the uniformity of the average frequency spectrum based on Shannon entropy [68]. It is defined as:

$$H_f = -\frac{1}{\log_2(N_f)} \sum_{f=1}^{N_f} s(f) \log_2 \left(s(f) \right). \tag{2.10}$$

Equation for Frequency Entropy (HFQ) [43], [67]

Where:

- $s(f) \in [0, 1]$ is the spectral probability mass function, obtained by averaging the spectrogram over time and normalizing over the frequency axis.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

- N_f is the total number of frequency bins.
- $H_f \in [0, 1]$ ensures normalised entropy.

Temporal Entropy (HTP)

The Temporal Entropy has been proposed, and utilised by Sueur et al [67]. Instead of frequency, temporal entropy analyses how energy varies over time [67]. The lower entropy means the sound intensity is stable (e.g., constant background noise such as wind, or rain). A high entropy means rapid intensity fluctuations (e.g., irregular bird calls). It is useful for detecting periodic sounds vs. sporadic vocalisations and identifying structured vs. chaotic sound environments [27].

The Temporal Entropy (H_t) quantifies the heterogeneity of an acoustic signal over time, based on Shannon entropy [68]. It is computed from the normalised amplitude envelope $a(n)$ as follows:

$$H_t = -\frac{1}{\log_2(N_t)} \sum_{n=1}^{N_t} a(n) \log_2 \left(a(n)\right). \tag{2.11}$$

Equation for Temporal Entropy (HTP) [43], [67]

Where:

- $a(n) \in [0, 1]$ is the normalised amplitude envelope at time step n.
- N_t is the total number of time samples.
- The normalization factor $\log_2(N_t)$ ensures $H_t \in [0, 1]$.

Combined Entropy (HHH)

The Combined Entropy has been proposed, and utilised by Sueur et al [67]. This index multiplies HFQ and HTP [67]. It provides an overall measure of sound variability in both frequency and time domains. While distinguishing between structured stable and noisy environments. It also provides a sense of stable sounds (low entropy) vs. dynamic complex soundscapes (high entropy) over time.

The overall Combined Entropy Index (H) is computed as the product of temporal and spectral entropy:

$$H = H_t \cdot H_f \tag{2.12}$$

Equation for Combined Entropy (HHH) [67]

This Combined Entropy Index reflects the acoustic complexity of the signal by capturing both temporal and spectral heterogeneity.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Acoustic Evenness Index (AEI)

The Acoustic Evenness Index has been developed by Bradfer-Lawrence et al. [53] and further researched by Pieretti et al., Villanueva-Rivera et al., and Luypaert et al. [21], [56], [57], [59], [61]. It evaluates how evenly sound energy is distributed across the entire spectrogram. A high AEI means all frequency bands are similarly active, indicating a balanced soundscape [21]. A low AEI suggests some frequency bands dominate, possibly due to human noise pollution. While differentiating urban vs. natural soundscapes, this metric Identifies imbalanced soundscapes (e.g., human activity dominating over wildlife, or specific animal species sounds on a particular frequency). This index ranges from 0 (perfect evenness) to values close to 1 (high inequality).

The Acoustic Evenness Index measures the evenness of acoustic activity across frequency bands and is based on the Gini [69] coefficient applied to the normalised energy distribution ρ_i obtained similarly to the ADI. Instead of a logarithmic weighting, AEI uses a pseudo-square form, reflecting inequality among bands.

Let N_i be the number of frequency bands and ρ_i the normalised energy in each band.

The Acoustic Evenness Index (AEI) is computed as:

$$\mathrm{AEI} = \frac{2\sum_{i=1}^{N_i} i \cdot \rho_i}{N_i \sum_{i=1}^{N_i} \rho_i} - \frac{N_i + 1}{N_i}. \tag{2.13}$$

Equation for Acoustic Evenness Index (AEI) [53]

Although AEI and ADI are derived from the same spectral data, they provide complementary information: ADI emphasizes diversity, while AEI quantifies evenness. Their values are typically negatively correlated.

Normalized Difference Soundscape Index (NDSI)

The goal of the Normalized Difference Soundscape Index is to estimate the level of anthropogenic disturbance on the soundscape by computing the ratio of human-generated (anthrophony) to biological (biophony) acoustic components found in field collected sound samples [70]. This is a composite index that compares: Natural sounds (e.g., birds, insects, wind) Anthropogenic (human-made) sounds (e.g., farming, construction, traffic, airplanes) [21]. A high NDSI indicates a healthy ecosystem dominated by natural sounds. A low NDSI suggests a high level of human noise pollution. It indicates the impact of urbanization on natural habitats.

Since Normalized Difference Soundscape Index measures the relative contribution of biophony and anthrophony within an acoustic environment. The NDSI can be calculated as follows:

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The Normalized Difference Soundscape Index (NDSI) is then calculated as:

$$\text{NDSI} = \frac{\text{BI} - \alpha}{\text{BI} + \alpha}, \quad \text{NDSI} \in [-1, 1]. \tag{2.14}$$

Equation for Normalized Difference Soundscape Index (NDSI) [43], [70]

Where the two spectral energy components are:

- BI is the acoustic energy from biological sources, typically computed over the frequency band $[f_{bio1}, f_{bio2}] = [2000, 8000]$ Hz.
- α is the acoustic energy from anthropogenic sources, computed over $[f_{anth1}, f_{anth2}] = [1000, 2000]$ Hz.

This normalised ratio ranges between -1 and 1. A value of NDSI $= 1$ indicates the complete absence of anthrophony, while NDSI $= -1$ implies no detectable biophony. Intermediate values reflect the relative dominance of each sound source, allowing assessments of soundscape composition over time.

2.3 Curated Dataset

The audio processing workflow receives as input-data the 57 soundscape recordings. These recordings are then preprocessed, transformed, and augmented through a series of signal processing steps as presented in Section 2.3.1. The result is a curated dataset composed of 4380 tensors per soundscape, each capturing a rich set of statistical and acoustic features. An overview of the resulting dataset is shown in Section 2.3.2.

2.3.1 Preprocessing Workflow

The audio processing workflow begins with 57 soundscape recordings, each of a duration of 30 minutes and sampled at 48000 Hz, resulting in a raw input shape of (1, 86400000) per soundscape. These recordings are then subsampled into 438 overlapping frames of 6 seconds each, with a 1.9-second overlap, and applied the Hanning function producing segments with a shape of (438, 288000). For each frame, a spectrogram is generated to represent the frequency content over time. Also an FFT is applied next, splitting each frame into 10 frequency bands of 1000 Hz bandwidth, sampled at 1000 samples per second, resulting in a shape of (438, 10, 6000). From these frequency bands, five basic statistical features are computed: mean, median, sum, maximum, and minimum, giving a feature matrix of shape (438, 10, 5). In parallel, 11 established Ecological Acoustic Indice (EAI), such as ACI, ACA, ADI, BET, MMM, NPP, HFQ, HTP, HHH, AEI, NDSI are calculated for each band and frame, yielding a shape of (438, 10, 11). Finally, the data is reshaped into a curated dataset of 4380 tensors per soundscape, where each tensor combines all extracted features into a flattened vector of 6016 values, resulting in a final

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

shape of (4380, 6016). The diagram of the data transformation for the preprocessing workflow is presented in the Figure 2.1.

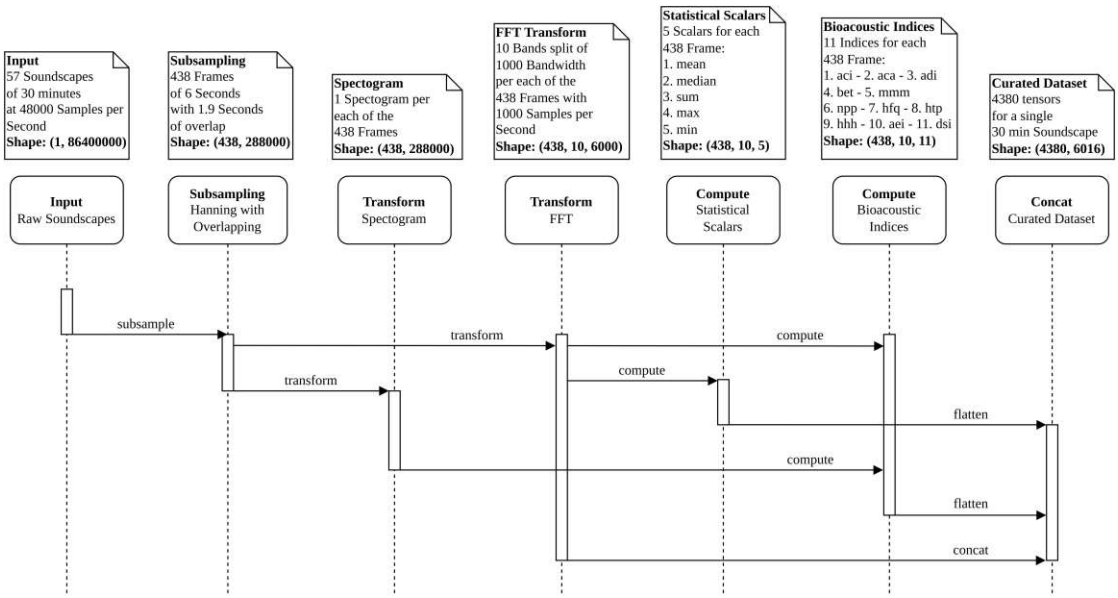


Figure 2.1: Data Transformation Sequence Diagram for Preprocessing

2.3.2 Curated Dataset Description

The curated dataset has a shape of 249,660 × 6,016. The golden labels are Plantation (96,360), Pasture (48,180), Natural Regeneration (91,980), and Reference Forest (13,140). Each entry has a duration of six (6) seconds with an overlap of 1.9 seconds resultant of applying the Hanning-window function with overlapping from Section 2.1.6 [48].

The filtered frames refers to the signal divided into 10 frequency bands as described in Section 2.1.9. The signal refers to the sounds only with Hanning-window applied and no-frequency-filtering as explained in Section 2.1.8. The description of the curated dataset is presented in the Table 2.1.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Table 2.1: Curated Dataset Description

Variable	Code	Description	Type	Values	Mean	Median
Region	reg	Geographic region ID	Category	[0 – 3]	–	–
Band	ban	Frequency band ID from 0 to 10000 Hz, split into 10 bands of 1000 Hz	Category	[0 – 10]	–	–
Mean Amplitude	men	Mean amplitude of the frame	Float	[-0.023 – 0.031]	-2.67×10^{-7}	5.81×10^{-7}
Median Amplitude	med	Median amplitude of the frame	Float	[-0.023 – 0.018]	-7.61×10^{-8}	3.45×10^{-8}
Aggregate Amplitude	sum	Total amplitude in the frame	Float	[-140.195 – 191.673]	-0.001	0.003
Maximum Amplitude	max	Highest amplitude value in the frame	Float	[0.01 – 27.3]	0.22	0.115
Acoustic Complexity Index	aci	Acoustic Complexity Index of the signal	Float	[-13029930 – 348855419]	2862.28	10.348
Alternative Acoustic Complexity Index	aca	Acoustic Evenness Index of the signal	Float	[75557 – 85746]	81058	81191
Acoustic Diversity Index	adi	Acoustic Evenness Index of the signal	Float	[0.808 – 2.824]	2.749	2.771
Beta Bioacoustic Index	bet	Beta Bioacoustic Index of the signal	Float	[2.54 – 76.975]	17.85	16.16
Temporal Median Index	mmm	Temporal Median Index of the signal	Float	[0 – 0.179]	0.002	0.001
Number of Peaks	npp	Total number of amplitude peaks in the frame	Integer	[43 – 73]	59	58
Frequency Entropy	hfq	Entropy across frequency bands	Float	[0.161 – 0.921]	0.735	0.759
Temporal Entropy	htp	Entropy over time in the frame	Float	[0.494 – 0.942]	0.895	0.908
Combined Entropy	hhh	Combined entropy	Float	[0.494 – 0.837]	0.895	0.908
Acoustic Evenness	aei	Acoustic Evenness Index of the signal	Float	[0.02 – 0.923]	0.319	0.305
Normalized Difference Soundscape Index	dsi	Normalized Difference Soundscape Index of the signal	Float	[0.0 – 1.0]	0.492	0.481
Vector	vec	Signal	Series	[1 x 6000]	–	–

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

2.3.3 Curated Dataset Overview

Based on the preprocessing workflow described in Section 2.3.1, we consolidate the curated dataset presented in Section 2.3.2, and subsequently generate visualisations to better understand the characteristics of each Reference Forest, Pasture, Natural Regeneration, and Plantation (Regions). The aim is to identify patterns and compare ecological Regions using the calculated EAIs as Aggregated scalar descriptors.

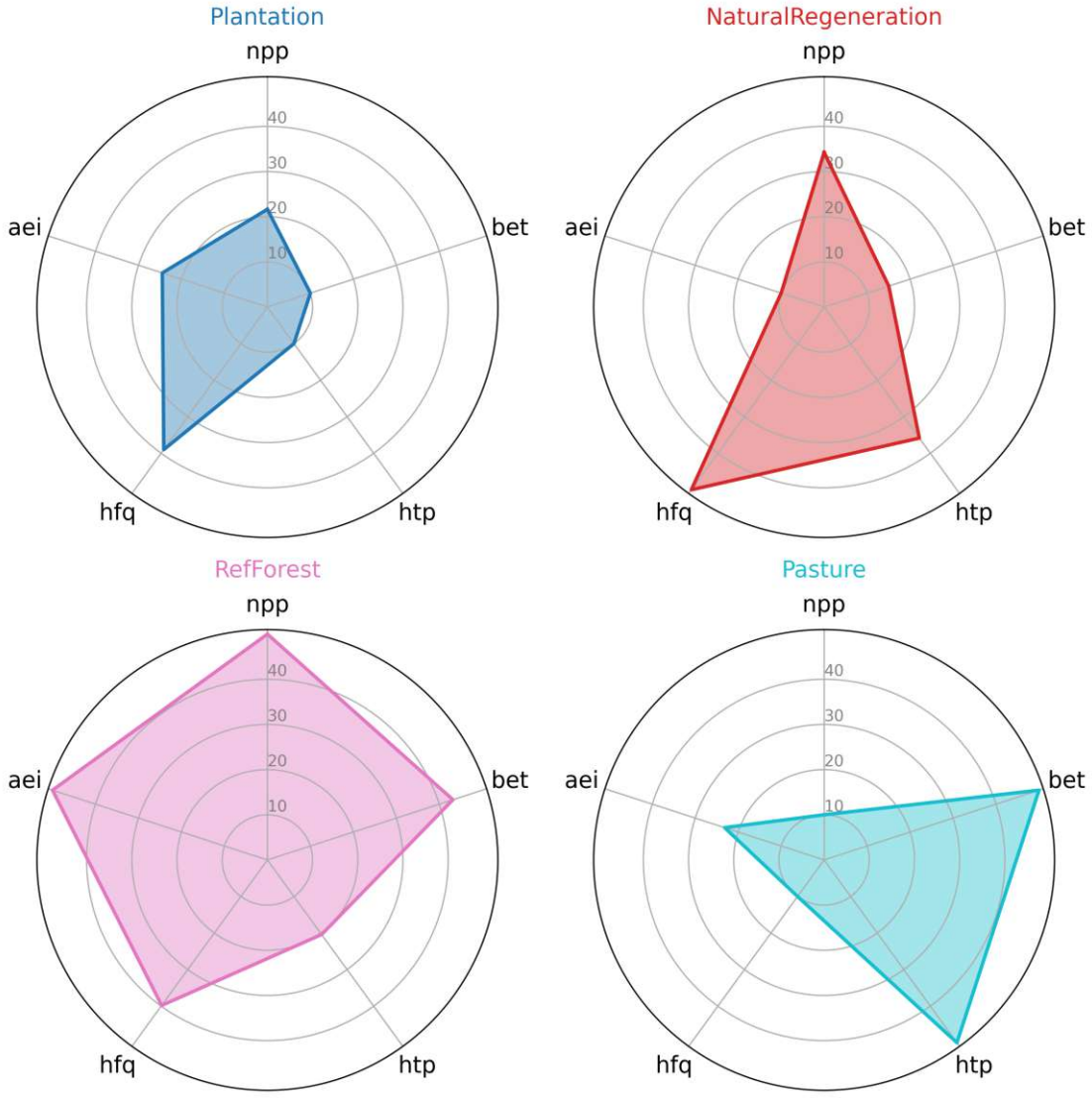


Figure 2.2: SpiderNet Chart of Ecological Acoustic Indices by Region

Figure 2.2 presents four (4) spidernet charts, one for each Reference Forest, Pasture, Natural Regeneration, and Plantation. Each chart displays normalised values, scaled between 10 and 50 units, for a subsample of five (5) selected EAIs: NPP, BET, HTP,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

HFQ, and AEI. The visualisation highlights distinct acoustic profiles across the Regions, illustrating how these indices vary among the different ecosystem types.

In Figure 2.2, Plantation (deep blue) shows minimal values for the BET and HTP indices, along with low values for AEI and NPP, while exhibiting moderate to high values for HFQ. In contrast, Pasture (aquamarine) presents the opposite trend, with the highest values for BET and HTP, minimal values for NPP and HFQ, and consistently low values for AEI. Natural Regeneration (orange) records the lowest AEI, low BET, moderate to high NPP and HTP, and the highest HFQ. Finally, the Reference Forest (magenta) exhibits consistently high values across most indices, with peak values for NPP and AEI, high HFQ and BET, but relatively low HTP.

Overall, these spidernet profiles provide a general outlook of the eco-acoustic variability across land-use types (Regions), supporting the hypothesis that specific EAIs can serve as ecological proxy-indicators for distinguishing ecosystems. This observation reinforces the motivation to investigate which subsets of EAIs most effectively enhance predictive performance in the subsequent chapters.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 3

Soundscapes Classification with Deep Learning

Chapter 3 introduces the FOREST family of Deep Learning (DL) architectures. Section 3.1 sets the theoretical foundation in Machine Learning and outlines the core methods used to tackle the supervised learning problem addressed in this thesis. The section also summarises Neural Networks as the underlying function class and motivates the use of DL-based models namely Convolutional Neural Networks and Long Short-Term Memorys, for classifying ecological audio recordings.

Section 3.2 presents, layer by layer, the technical specifications of the FOREST hybrid DL architectures: SeqNet-CNN-LSTM, SeqNet-LSTM-CNN, and ParaNet-CNN-LSTM. For rigorous comparison and benchmarking, Section 3.2.4 introduces the baseline models employed for this purpose: Simple-SVM, Simple-CNN, Simple-LSTM, and ResNet34.

Section 3.3 describes the training configuration applied to all models, including the dataset partitioning, learning criteria, and serialisation strategies, carried out using the computational power and GPU resources of two HPC Systems: Izar [71], operated by SCITAS at the EPFL, and Vienna Scientific Cluster (VSC)-5 [72], provided by the Austrian Scientific Computing (ASC) at the TU Wien.

Finally, Section 3.4 defines the evaluation methodology and presents the preliminary results of the empirical experiments presented in Section 4.1. The models are assessed using the curated dataset from Section 2.3, applying Accuracy as the main performance metric. The baseline models are compared against the proposed FOREST variants, providing an initial overview of their effectiveness for ecological soundscape classification.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

3.1 Deep Learning Basics

This section introduces the core concepts of Machine Learning and Neural Network, followed by a brief overview of Deep Learning. It highlights the key characteristics of the foundational model architectures utilised in this master thesis, namely Convolutional Neural Networks, and Long Short-Term Memory models.

3.1.1 Machine Learning

Machine Learning (ML) is a collection of techniques that enable the automatic detection of patterns in data, which can then be leveraged to make predictions about future or unseen data [73]. A well known definition by Mitchell states that "a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" [74]. According to this definition, and particularly in the context of this thesis, the task T is to recognise and classify acoustic characteristics in audio recordings, the performance measure P corresponds to the percentage of correctly classified examples (Accuracy), and the experience E is given by a labelled dataset of annotated Regions.

Formally, a supervised learning problem is defined by a training set $D = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i denotes the input and y_i the corresponding target label. The aim of the learning algorithm is to approximate a function $f : X \to Y$ that maps inputs to outputs. If y is continuous, the task is referred to as regression, whereas if y is categorical, the task is classification [73], [74]. Beyond classification and regression, Machine Learning can also be of type unsupervised learning and reinforcement learning [75]. In unsupervised learning, no labels are provided, and the goal is to discover latent structures in the data, such as clusters or lower-dimensional representations. Reinforcement learning, by contrast, involves agents interacting with an environment and improving their behaviour based on delayed feedback in the form of rewards or penalties.

A core requirement in Machine Learning is generalisation, meaning that models are expected to perform well not only on the training data but also on unseen examples. Performance is therefore typically measured on a test set that is distinct from the training data [75]. In addition, a validation set is often employed during training to fine-tune hyperparameters and mitigate overfitting [76]. Accuracy, defined as the proportion of correctly classified examples, is a common metric for classification tasks, whereas the error rate captures the proportion of misclassified instances [77].

Despite its effectiveness, supervised learning has important limitations. The performance of modern Deep Learning (DL) models strongly depends on the availability of large labelled datasets, sometimes requiring thousands of annotated examples per category and millions of examples to reach human-level capabilities [75]. However, producing such datasets is costly and time-consuming, since labelling often requires expert knowledge and must be maintained in time. Several approaches have been proposed to mitigate this issue. Active learning aims to identify the most informative data points for annotation,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

thereby reducing labelling effort. Transfer learning leverages knowledge from related tasks to improve performance on a target task using limited labelled data [78], [79].

In this thesis, we focus on supervised learning, particularly for classification tasks. The models are trained on annotated bioacoustic datasets, with training and validation subsets supporting parameter learning and hyperparameters tuning, respectively.

3.1.2 Neural Networks

Neural Networks (NNs) are a type of Machine Learning algorithm inspired by the structure of the human brain [76], [80]–[82]. A neuron is a computational unit that receives multiple inputs, computes a weighted sum, adds a bias, and applies a non-linear activation function to produce an output. Neurons are organised into layers, and information flows from the input layer through one or more hidden layers to the output layer. Layers other than the input and output are called hidden layers, as their activations are not directly observable. The number of layers defines the depth of the network, while the dimensionality of the hidden layers defines its width [75], [83].

Feedforward Neural Networks (FNNs) are the simplest type of NNs, where information flows in one direction without feedback loops. The function computed by an NN is a multivariate non-linear mapping from inputs to outputs, where the output of each neuron is determined by its inputs, weights, and biases [84]. Training consists of minimising a loss function that measures the discrepancy between predictions and actual labels [76].

Parameters such as weights and biases are learned during training, while hyperparameters such as learning rate, batch-size, and choice of activation function (e.g., ReLU or sigmoid functions) must be set previous the training stage.

3.1.3 Deep Learning

Deep Learning (DL) extends traditional NNs by employing multiple hidden layers, which enables the extraction of complex hierarchical features from raw data [75], [83]. Unlike shallow ML methods that often rely on hand-crafted features, DL models jointly learn representations and perform classification within a single architecture. This capability has proven particularly effective for high-dimensional data such as images and audio signals, and its impact on science and technology was recognised with the 2024 Nobel Prize in Physics [85], [86].

Activation functions play a central role in Deep Learning. The Rectified Linear Unit (ReLU), defined as $p(x) = \max(x, 0)$, is widely used because of its simplicity and favourable optimisation properties [75], [76]. Other functions, such as the logistic sigmoid $S(x) = \frac{1}{1+e^{-x}}$, are often used when outputs must be interpreted as probabilities [74]. This optimisation is often carried out using Stochastic Gradient Descent (SGD) or its variants, where model parameters are iteratively updated in the direction that minimises the loss-function [74], [76], [87].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Backpropagation is the algorithm used to compute these gradients efficiently by propagating the error backward through the network [76]. Formally, for each parameter $\phi_i \in \Phi$, the update-rule is given by the following equation.

$$\Delta\phi_i = -\eta \frac{\partial L}{\partial \phi_i}. \tag{3.1}$$

Equation for Gradient Descent update-rule [74]

Where η is the learning-rate that controls the step size of the update [74]. The negative sign ensures that parameters are moved in the direction that decreases the loss-function. Training proceeds iteratively over the dataset in epochs, where each epoch corresponds to one complete pass through the training data.

3.1.4 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are partially connected NNs designed for pattern recognition tasks [88]. Unlike fully connected networks, CNNs exploit three principles to improve efficiency and generalisation: local receptive fields, weight sharing, and pooling. Convolutional layers apply learnable filters (kernels) across small regions of the input, generating feature maps that capture local patterns. Because the same kernel is applied across the entire input, CNNs are able to detect translation-invariant features while greatly reducing the number of parameters [75], [83].

Learning in CNNs is hierarchical, early layers detect simple structures such as edges, intermediate layers capture higher-order patterns, and deeper layers combine these into abstract semantic features [89]. Pooling layers further downsample the feature maps, improving robustness to small shifts and distortions while reducing the computational complexity of the model [90]. At the final stage, one or more Fully Connected (FC) layers integrate the learned representations and perform the classification task [91]. Additional regularisation techniques, such as dropout, are often employed to prevent overfitting by randomly omitting activations during training [92].

CNNs have achieved state-of-the-art performance across a wide range of domains, including image classification, object detection, semantic segmentation, and audio-based classification [75], [77], [91]–[93]. Their ability to automatically extract and generalise local features has made them one of the most influential architectures in modern Deep Learning. Beyond computer vision, recent research has demonstrated state-of-the-art results for time-series forecasting using DL-based models [28], [29], [94]–[101], and even for soundscape classification [6], [26], [30], [102].

3.1.5 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a class of NN specifically designed to handle sequential data by incorporating recurrent connections between hidden units [75]. Unlike

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

feed-forward networks, which assume independence between inputs, RNNs introduce temporal dependencies by allowing information to persist across time steps. This is achieved through feedback loops in which the hidden state at time t depends not only on the current input but also on the hidden state at $t-1$, thereby introducing memory into the network [103], [104]. In graph terminology, RNNs can be unfolded across time, forming a directed acyclic graph where each node corresponds to a time step [75].

Formally, RNN process an input sequence $x = (x_1, \ldots, x_T)$ by iteratively updating a hidden state $h_t = f(Wx_t + Uh_{t-1} + b)$, where W and U are weight matrices and f is a non-linear activation function. The output sequence $o = (o_1, \ldots, o_T)$ is computed as $o_t = g(Vh_t)$ with output weights V. This recurrent structure enables RNNs to model both short- and long-term temporal dependencies, making them well suited for tasks such as speech recognition, text processing, and video analysis [105]. In the domain of audio analysis, signals can be represented as sequences of frames or feature vectors, and directly processed with RNNs [105].

Several extensions of RNNs, such as Convolutional Recurrent Neural Networks (CRNN), combine convolutional layers to capture local spatial features with recurrent layers to model longer temporal dependencies [91], [106]. Other variations, such as the Gated Recurrent Unit (GRU), and Long Short-Term Memory (LSTM), were specifically designed to overcome the vanishing and exploding gradient problems inherent to RNNs [107]. These gating-based mechanisms have proven particularly effective for sequential data in general, and audio modelling in particular [108].

3.1.6 Long Short-Term Memory

Long Short-Term Memorys (LSTMs) are a specialised type of RNN, first proposed by Dr. Sepp Hochreiter in his master thesis at Technische Universität München (TUM) in 1991 [109] and later formalised in a publication at MIT Press in 1997 [107]. They were explicitly developed to address the vanishing and exploding gradient problems of conventional RNNs, which limits the modelling of long-term dependencies in sequential data [104], [107], [109]. To overcome these limitations, LSTMs introduce memory cells and a gating mechanism (input, output, and forget gates) that regulate the flow of information across time-steps, allowing selective retention, updating, or discarding of information to effectively learn the long-range temporal data-dependencies [75], [110].

Formally, an LSTM cell maintains an internal state c_t in addition to the hidden state h_t. The gates, implemented as sigmoid activations, control how much of the previous state is kept or updated. The input gate regulates the incorporation of new information, the forget gate determines what information to discard, and the output gate controls the exposure of the internal state. This structure mitigates gradient vanishing, allowing the network to back-propagate errors across longer sequences effectively [107].

LSTMs have been widely applied to time-series modelling and forecasting. In audio-related applications, they are particularly effective at capturing temporal dynamics across

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

feature frames such as spectrograms, and have been implemented in tasks such as speech recognition, machine translation, and audio classification [108], [111].

Recent advances from the Johannes Kepler Universität Linz (JKU) have led to the development of Extended Long Short-Term Memory (xLSTM) [112], which extend the traditional gating structure of LSTMs by integrating attention mechanisms [113]. Through innovations such as exponential gating for long-sequence memory control and matrix memory for enhanced storage and parallelisation, xLSTMs achieve higher accuracy and efficiency while avoiding the quadratic complexity of standard attention [112]. As a result, LSTM-based models have not only outperformed transformer architectures in several tasks [112], [114], [115], but also demonstrated strong performance in NLP-based applications, where sequential dependencies are crucial.

3.2 Model Architectures

In the context of this master thesis, the primary aim is to introduce the FOREST within the use case of multi-class classification of soundscapes under the Payment for Environmental Services (PSA) program. This involves classifying time-series data from soundscape recordings by assigning categorical labels to input sequences based on their temporal dynamics, underlying dependencies, and extracted features. Achieving accurate classification requires models that can capture both local structures and long-range sequential information. To this end, Convolutional Neural Networks (CNNs) are employed for their ability to extract meaningful features from spectrograms and other image-like representations, while Long Short-Term Memorys (LSTMs), as a specialised type of RNNs, are integrated to model dependencies across time. Together, these architectures form hybrid Deep Learning models capable of learning complementary spatial and sequential characteristics of ecological-acoustic data, which provide the methodological foundation for this thesis [83].

The proposed approach integrates CNNs and LSTM models, applied in both sequential and parallel configurations, to carry out the audio classification task similar to [110], [116]–[132]. Each of these DL model architectures bring distinct strengths to execute the classification task. The CNNs are highly effective in identifying short-term, localised patterns through their sliding filter mechanism, also known as convolution (see Section 3.2.1) [80], [82], [84], [89], [114], [133], [134].

Given their complementary capabilities, integrating CNN and LSTM into hybrid models has gained popularity in the domain of time series analysis. These hybrid architectures aim to exploit the strengths of both networks, often resulting in enhanced performance compared to using either only one model in isolation. Recent studies have explored this integration, proposing various strategies for structuring the flow of information between CNN and LSTM layers in the context of time-series analysis, such as [29], [95], [101], [132], [135]–[145]. However, only a few studies have applied these hybrid DL architectures within the audio classification domain [131], [132], [146], and to date, there is no evidence that this approach has been explored directly in the bioacoustic domain.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Hybrid CNN - LSTM architectures can be particularly well-suited for bioacoustic audio classification not just because they effectively combine the complementary strengths of both models. Furthermore, CNN have proven performance at extracting detailed spectral features from stationary or quasi-stationary sound patterns (e.g. animal observations) while LSTM are capable of understand the temporal dynamics of non-stationary or evolving acoustic events (e.g. sequences of vocalisations, and even sounds in motion). These bioacoustic records often exhibit a mix of these spatial and temporal structures, embedded in noisy, overlapping, and highly variable environments. Therefore, our proposed hybrid models aim to tackle these complexities by jointly capturing localised spectral sounds within temporal dependencies, leading to improved robustness, classification Accuracy (AC), and even generalisation across diverse and real-world datasets.

In this thesis, we propose a hybrid ensemble strategy that combines DL model architectures for the first time within the environmental soundscapes analysis. Specifically, we experiment with three CNN - LSTM hybrid architectures: SeqNet-CNN-LSTM (Section 3.2.1), SeqNet-LSTM-CNN (Section 3.2.2), and ParaNet-CNN-LSTM (Section 3.2.3).

3.2.1 Sequential Network CNN - LSTM (SeqNet-CNN-LSTM)

Le et al. (2019) [137] were among the first to experiment with this hybrid DL model architecture. In their study, the proposed model extracted spatial and temporal features to effectively predict energy consumption using the Individual Household Electric Power Consumption (IHEPC) dataset [147]. Although the approach demonstrated success in forecasting electricity consumption, further improvements in predictive performance on the IHEPC dataset were necessary. Therefore, Le at al. introduced an improved model, named as Electric Energy Consumption Prediction model utilising the combination of Convolutional Neural Network and Bi-directional Long Short-Term Memory (EECP-CBL) [137]. The experimental results confirmed that the proposed model outperformed existing approaches in predicting household energy consumption on the IHEPC dataset.

These promising results in outperforming state-of-the-art models for energy consumption forecasting inspired subsequent research in the time-series domain and led to the widespread adoption of this hybrid architecture. Therefore, this is one of the most prominent state-of-the-art configurations among those presented in this thesis with proven results used for time series analysis [29] in electrical consumption [137], finance prediction [138], air temperature prediction [95], soil carbon concentration prediction [139], geographical location of marine vessels [140], influenza forecasting [101], mine water inflow [141], air quality forecasting [143], acoustic scene recognition [131], [132], [144], [146], [148], and medical [142], [145] domains.

SeqNet-CNN-LSTM Architecture

The Sequential Network CNN - LSTM (SeqNet-CNN-LSTM) model is composed by convolutional and recurrent layers as follow.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The architecture begins with two 1Dimensional (1D) convolutional layers: The first one applies 64-filters to extract low-level local features from the input-vector, while the second uses 128-filters to capture more abstract and hierarchical spectral representations. Each convolutional layer is followed by a Rectified Linear Unit (ReLU) activation and a Maximum (MAX)-pooling operations, which introduce non-linearity, improve generalisation, and emphases relevant features while reducing the temporal dependencies of the signal.

The resulting feature tensor is then fed into a multi-layer LSTM network with the aim to learn about the sequential dependencies and retain contextual information over time. Which is the relevant importance for capturing the dynamic structure of bioacoustic signals. Finally, a fully connected layer condenses the LSTM output to the target classes.

Here, we give a brief explanation of each of the layers:

Layer 1: Input Permutation

The SeqNet-CNN-LSTM model is designed to process 1D-input-tensors structured as (N, T, F), where N denotes the batch size, T the number of time steps, and F the number of features (also known as channels in this context) per time step. However, the internal CNN layer in PyTorch expects input in the format (N, F, T), with the feature dimension as the channel axis. To satisfy this requirement, the model applies a permutation operation that transposes the input from (N, T, F) to (N, F, T), enabling temporal convolution across the sequence length [149].

Layer 2: 1st 1D-Convolution

The 1D-Convolution operates by sliding a small filter across the vector to detect meaningful patterns, such as changes or repetitions in the time-series. Each filter is trained to highlight specific features, producing a new vector that marks where those features appear. This helps the model focus on relevant structures—like sound events from the input audio signal, while reducing less important or noisy information [134].

The received tensor from Layer 1 is fed into the 1D-Convolution operation [150].

Operation: $\texttt{Conv1d}(\text{in} = F, \text{out} = 64, \text{kernel\_size} = 3, \text{stride} = 1, \text{pad} = 1)$

This operation applies a 1D-Convolution over the input tensor composed of several input planes. In the simplest case, the output value of the layer with input size (N, C_{in}, L) and output size $(N, C_{\text{out}}, L_{\text{out}})$. Where padding equals 1 (which is the value added at both sides of each tensor), N is the batch size, C denotes the number of channels, and L is the length of the input sequence [150].

The equation can be described as follow:

$$\text{out}(N_i, C_{\text{out}_j}) = \text{bias}(C_{\text{out}_j}) + \sum_{k=0}^{C_{\text{in}}-1} \text{weight}(C_{\text{out}_j}, k) \star \text{input}(N_i, k). \tag{3.2}$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

1Dimensional (1D) Convolution equation [150]

Where $\star$ is the valid cross-correlation operator.

Activation: Rectified Linear Unit

The Rectified Linear Unit (ReLU) outputs the input directly if it is positive; otherwise, it outputs zero [151]. This activation function is applied to each element of the output from Layer-1. The ReLU can be defined as follow:

$$\text{ReLU}(x) = (x)^+ = \max(0, x). \tag{3.3}$$

Rectified Linear Unit (ReLU) equation [151].

Pooling: `MaxPool1d`$(\text{kernel\_size} = 2, \text{stride} = 2)$

The 1D MAX Pooling operation reduces the dimensionality of a sequence by selecting the Maximum (MAX) value within fixed-size window as it slides across the input. This process helps retain the most prominent features while discarding less relevant details, making the model more robust to noise and improving computational efficiency. For this specific audio task, it highlights strong sound observations and compresses the temporal resolution of the signal [152].

The `MaxPool1d` function is applied over the given signal composed of several input planes received from the ReLU. It receives an input tensor of shape (N, C, L) and an outputs a tensor of shape (N, C, L_{out}). The function is defined as as follow:

$$\text{out}(N_i, C_j, k) = \max_{m=0,\ldots,\text{kernel\_size}-1} \text{input}(N_i, C_j, \text{stride} \times k + m). \tag{3.4}$$

1Dimensional (1D) Max Pooling equation [152]

Where N represents the batch size, C is the number of channels, L is the length of the input sequence. The stride is the pooling step size, and kernel_size is the width of the pooling window [152].

Output Shape: $(N, 64, T/2)$

After the 1-Dimensional Max Pooling, the length of the sequence is reduced by keeping only the most important features (Maximum value) in each fixed-size window stride per stride. In this operation, we received an input of size 128, and an delivered an output of size 64. It helps in downsampling, which reduces the computation and provides spatial invariance.

Layer 3: 2nd 1D-Convolution

The structural pattern of the first 1D-Convolutional layer is repeated. However, the input-output dimensions of the tensors change, as the output of the 1st 1D-Convolutional layer (Layer 2) becomes the input to the 2nd 1D-Convolutional layer (Layer 3). The

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

purpose of this 2nd convolutional operation is to extract higher-level features from the previously processed representations, while further reducing the temporal dimension by half. As in Layer 2, Layer 3 applies the same sequence of operations (convolution, activation, and pooling) with updated input-output dimensions, as follows:

Operation: `Conv1d`$(64, 128, \text{kernel\_size} = 3, \text{stride} = 1, \text{pad} = 1)$.

Activation: ReLU.

Pooling: `MaxPool1d`$(\text{kernel\_size} = 2, \text{stride} = 2)$.

Output Shape: $(N, 128, T/4)$

Layer 4: Permutation for LSTM

The output tensor from the CNN component has the shape $(N, 128, T/4)$, where 128 represents the number of extracted feature channels, and $T/4$ reflects the reduced temporal length after two 1D MAX Pooling operations. However, the PyTorch LSTM component [153] expects input tensors in the format (N, T, F), where T is the sequence length and F the number of features at each time step. To meet this requirement, a permutation operation is applied to transpose the tensor from $(N, 128, T/4)$ to $(N, T/4, 128)$, enabling the LSTM to model temporal dependencies across the processed feature sequence [149].

Layer 5: Long Short-Term Memory (LSTM)

This intermediate PyTorch LSTM module [153] within the SeqNet-CNN-LSTM architecture is responsible for modelling temporal dependencies in the input sequence by applying a multi-layer LSTM recurrent neural network, where each layer processes the input over time by computing a sequence of gated transformations that retain long-term contextual information [107], [109], [110].

The Long Short-Term Memory (LSTM) network is a type of Recurrent Neural Network (RNN) designed to model temporal sequences and long-term dependencies. It achieves this by incorporating memory cells and gating mechanisms to regulate the flow of information [107], [109], [110].

For each time step t, the LSTM module performs the following computations:

$$
\begin{aligned}
i_t &= \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}),\\
f_t &= \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}),\\
g_t &= \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}),\\
o_t &= \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}),\\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t,\\
h_t &= o_t \odot \tanh(c_t).
\end{aligned}
\tag{3.5}
$$

PyTorch LSTM implementation workflow [153].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Where x_t denotes the input at time t, h_t is the hidden state, c_t is the cell state, and i_t, f_t, g_t, and o_t represent the input, forget, candidate, and output gates, respectively. The σ function is the sigmoid activation, and $\odot$ denotes the Hadamard (element-wise) product [153].

In a multilayer LSTM, the input to layer l ($x_t^{(l)}$) is typically the hidden state $h_t^{(l-1)}$ from the previous layer, optionally regularised by dropout[92], [153]:

$$x_t^{(l)} = \delta_t^{(l-1)} \cdot h_t^{(l-1)}. \tag{3.6}$$

LSTM Dropout function [92], [153].

Where $\delta_t^{(l-1)}$ is a Bernoulli random variable that equals 0 with dropout probability p. If a projection layer is used (when `proj_size > 0`), the LSTM modifies the hidden state h_t using a learnable linear transformation [153]:

$$h_t = W_{hr}\tilde{h}_t. \tag{3.7}$$

LSTM Linear Transformation function [153].

Where $\tilde{h}_t$ is the output of the original LSTM cell and W_{hr} is the projection matrix. Finally, this operation reduces the dimensionality of the output and enables more efficient training and inference [153].

In our proposed architecture, the LSTM is configured with an `input_size` equal to the number of features per time step (i.e., `X_train.shape[-1]`), a `hidden_size` of 128, and `num_layers` set to 2. The `batch_first=True` flag ensures that input tensors follow the shape (N, T, F), where N is the batch size, T the sequence length (i.e., time steps), and F the number of features per time step.

Operation: $\texttt{LSTM}(input\_size, \text{hidden\_size}, \text{num\_layers}, \text{batch\_first} = True)$.

Intermedium Output: The hidden states for each time step. However, only the last hidden state (i.e., h_t at the final time step) is passed to the classifier.

Final Output Shape: $(N, \text{hidden\_size} = 128)$.

Layer 6: Fully Connected (FC) Linear Layer for Final Classification

The final step in the SeqNet-CNN-LSTM architecture involves a PyTorch Fully Connected (FC) Linear Layer module, which receives the last hidden state of the LSTM with shape (N, H), where N is the batch size and H is the hidden size. This layer performs an affine transformation of the form $y = xW^\top + b$, mapping the high-level temporal features into a new space of dimension C, corresponding to the number of output classes [154]. The resulting output tensor of shape (N, C) contains unnormalised

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

class scores (also known as logits), which can then be passed to a soft-max function or used directly for loss computation during training process (See Section 3.3) [154].

A Fully Connected (FC) (also known as dense) Linear Layer applies an affine transformation to the input data described as [154]:

$$\mathbf{y} = \mathbf{x}\mathbf{A}^\top + \mathbf{b}. \tag{3.8}$$

Fully Connected (FC) Linear Layer function [154].

Where $\mathbf{x} \in \mathbb{R}^{*,H_{\text{in}}}$ represents the input tensor with $H_{\text{in}} = \texttt{in\_features}$ being the dimensionality of each input sample. The matrix $\mathbf{A} \in \mathbb{R}^{H_{\text{out}} \times H_{\text{in}}}$ is a learnable weight matrix, and $\mathbf{b} \in \mathbb{R}^{H_{\text{out}}}$ is a learnable bias vector. The output tensor $\mathbf{y} \in \mathbb{R}^{*,H_{\text{out}}}$ has a shape where $H_{\text{out}} = \texttt{out\_features}$ denotes the size of each output sample [154].

Operation: $\texttt{Linear}(\text{hidden\_size} = 128, \text{num\_classes} = 4)$

Final Output: Logits of shape $(N, \text{num\_classes})$

This Layer takes the final hidden state vector (of dimension hidden_size) and maps it into a vector of length num_classes. Where each output dimension corresponds to a singular golden label for the classification task. The FC layer translates the final LSTM summary into a prediction over classes.

3.2.2 Sequential Network LSTM - CNN (SeqNet-LSTM-CNN)

The hybrid architecture combining CNN and LSTM has gained significant attention across multiple time-series domains due to its ability to effectively model both temporal and spatial dependencies in sequential data [29], [95], [101], [131], [132], [135]–[146].

However, the ordering of these components is not arbitrary. It is rooted in the complementary nature of their respective strengths. As LSTMs are explicitly designed to capture long-range temporal dependencies through gated memory mechanisms, which makes them well-suited for sequence modelling tasks. By processing the input sequence first with an LSTM, the model learns a temporally contextualised representation that retains dynamic trends and correlations across time [135], [136]. When this representation is subsequently passed to a CNN, the convolutional layers are better equipped to extract localised patterns (e.g., peaks, transitions, repetitions) that are meaningful within the temporal context already modelled by the LSTM.

In contrast, reversing this order, as in SeqNet-CNN-LSTM models, means the initial convolution layers extract spatial patterns without access to long-range temporal structure. These pooled features, while potentially useful, can lose the original temporal sequencing crucial for many time-dependent patterns [135], [136]. This limits the ability

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

of downstream LSTMs to infer time-based relationships that may potentially be already vanished or abstracted away by prior spatial operations.

Although hybrid Deep Learning models combining recurrent and convolutional layers have gained popularity in time-series analysis, there is limited evidence in the literature supporting the superiority of LSTM-CNN architectures over the more commonly adopted CNN-LSTM models. The existing studies implementing the LSTM-CNN ordering have primarily focused on structured time-series data in domains such as banking [136], and electrical energy forecasting [135]. These domains typically involve clearly defined and regular sequential patterns. However, their applicability to unstructured or semi-structured data types such as environmental audio records has not yet been extensively validated.

In this thesis, we propose to evaluate the performance of this hybrid Sequential Network LSTM - CNN (SeqNet-LSTM-CNN) architecture on bioacoustic environmental audio recordings to empirically test this hypothesis and determine its effectiveness for spatio-temporal feature learning in acoustic classification tasks.

SeqNet-LSTM-CNN Architecture

The Sequential Network LSTM - CNN (SeqNet-LSTM-CNN) model follows a hybrid architecture that integrates recurrent and convolutional layers. The network starts with a stacked multi-layer Long Short-Term Memory (LSTM), which processes the input signal to model long-term dependencies and contextual information over time.

The temporally encoded output from the LSTM is then reshaped and passed into a Convolutional Neural Network (CNN) module. The convolutional module begins with a 1D-convolutional layer with 64 filters, followed by a second with 128 filters, both using a kernel size of 3. These layers specialise in extracting localised spectral patterns and spatial features from the previous LSTM-time-contextualised representation. Each convolution is followed by a ReLU activation function and a 1D Max Pooling operation to reduce dimensionality and enhance discriminative feature learning.

To transition from convolutional feature vectors to class predictions, the output is flattened and processed through a FC Linear function. This last operation consists of a lazy linear layer that adapts to the input dimension, followed by another ReLU activation function and a final linear layer that maps the learned feature representation to the target class-space.

Overall, this architecture leverages the strengths of LSTMs in modelling sequence-level patterns before extracting local feature hierarchies with CNNs, providing a structured and effective spatio-temporal learning pipeline.

Here, a brief explanation of each layer in the proposed SeqNet-LSTM-CNN architecture is provided below, with an emphasis on the fact that the DL components are generally the same as those used in the SeqNet-CNN-LSTM model as follow.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Layer 1: Long Short-Term Memory (LSTM)

The first component of the SeqNet-LSTM-CNN model is a stacked LSTM network, which belongs to the family of Recurrent Neural Networks (RNNs). Its primary task is to capture temporal dependencies within the sequential input data, making it particularly well-suited for the soundscape classification task. The configuration, input format, and output structure of this component are identical to those used in the SeqNet-CNN-LSTM model (see Layer 5 in 3.2.1). However, unlike the initial CNN-based variant, this model does not require a preceding permutation operation, as the LSTM layer in PyTorch [153] can accept input tensors in the shape (N, L, F). Where N is the batch size, L is the sequence length, and F is the number of features per time step. This compatibility is ensured by setting the `batch_first=True` argument during instantiation, allowing the SeqNet-LSTM-CNN to directly process the input sequences without reordering the tensor dimensions.

Layer 2: Permutation for CNN module

The LSTM output is permuted identically as in the Layer 1 of the SeqNet-CNN-LSTM to match the expected input format of the 1-Dimensional Convolution operation (See Layer 1 in Section 3.2.1).

Layers 3-4: 1st & 2nd 1D-Convolutions

Following the LSTM layer, and its output permutation to match the required input-shape for the 1D-Convolution operations. Here, the tensor is fed into a two layers of of 1D-Convolutional [150] modules identically as the modules implemented in the SeqNet-CNN-LSTM model (See Layers 2 & 3 in Section 3.2.1). The difference here is tat these input tensors operate over the temporal axis of the LSTM hidden representations.

The 1st 1D-Convolutional module begins with a `Conv1d` layer that takes as input 128 channels—corresponding to the LSTM hidden size—and applies 64 learnable filters with a kernel size of 3, a stride of 1, and padding of 1. This configuration preserves the sequence length while enabling the extraction of local temporal patterns. The convolution is immediately followed by a ReLU activation function [151], introducing non-linearity, and a 1D MAX Poling operations [152] with a kernel size and stride equals 2, effectively downsampling the temporal resolution by a factor of 2.

The 2nd 1D-Convolutional module mirrors the structure of the first, now operating on 64-channel input (after the downsampling of the temporal resolution by a factor of 2), and applying 128 filters. Once again, the same convolutional parameters are used, followed by the same workflow: A ReLU activation and 1D MAX Pooling operations.

These convolutional modules together act as a hierarchical feature extractor, reducing temporal dimensionality (1st 1D-Convolution) while expanding the representational depth (2nd 1D-Convolution).

After both 1D-Convolutional modules, the temporal dimension of the output is reduced to one quarter of the original LSTM sequence length, and the resulting tensor has the shape $\mathbb{R}^{N\times 128\times (L/4)}$, where N is the batch size and L is the original vector length.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

This resultant reshaped tensor captures high-level, temporally compressed, and spatially rich patterns from the input tensors.

Layer 5: Flattening

The output tensor from Layer 4 is then flattened [155] into a 2-Dimensional (2D) shape (N, F) where $F = 128$ is total number of features after the convolutional layers, preparing it for Fully Connected (FC) classification layers.

Layer 6: Fully Connected Lazy Linear Layer for Pre-Classification

In this FC Lazy-Linear [156] layer, the number of input features is inferred automatically from the last dimension of the input tensor at runtime. This allows the layer to adapt flexibly to the preceding layer's output and project it into a 256-dimensional space through a standard linear transformation.

Operation: `LazyLinear`(out_features $= 256$)

Activation: Rectified Linear Unit (ReLU) (See Layer 2 in Section 3.2.1).

Layer 7: Fully Connected (FC) Linear Layer for Final Classification

This layer takes the final hidden state vector (of dimension hidden_size equals 256) and maps it into a vector of length num_classes [154]. Where each output dimension corresponds to a singular label as result for the classification task exactly the same as implemented in the SeqNet-CNN-LSTM (See Layer 6 in Section 3.2.1.

Operation: `Linear`(hidden_size $= 256$, num_classes $= 4$)

Output: Logits of shape $(N, \text{num\_classes} = 4)$

In the end of this workflow, the last FC layer performs the final classification by projecting the 256-dimensional feature vector into a vector of class logits, each representing one of the target categories among the target Regions (Reference Forest, Pasture, Natural Regeneration, and Plantation).

3.2.3 Parallel Network CNN - LSTM (ParaNet-CNN-LSTM)

Farsie et al. (2021) [135] was one of the first to explore the parallel hybrid CNN - LSTM model architecture in the domain of Short-Term Load Forecasting (STLF). Their task was to maintain the balance between electricity supply and demand within smart grids.

The researchers propose a hybrid DL model named Parallel LSTM - CNN Network (PLC-Net) that brings a novel variation to the use of LSTMs and CNNs. Unlike conventional architectures that combine these layers sequentially, where one network processes the input before passing it to the next component. PLCNet executes the LSTM and CNN modules in parallel. This parallel approach enables each subnetwork to independently extract the features enabling the LSTMs to uncover temporal patterns across time, while

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

CNNs to capture local or structural patterns within the input data. The two processed outputs are then combined at a later stage in the model.

This design drives to several advantages: First, it avoids the risk of losing important temporal or spatial features that might be suppressed when processed sequentially. For instance, if temporal dependencies are altered before spatial patterns are extracted, or vice versa, critical patterns may be missed. The parallel configuration allows both types of features to be preserved and interpreted on their own terms, which contributes to more accurate learning.

Farsie et al. [135] evaluated the model using two real-world datasets [157]. The first consisted of hourly electricity consumption records from Johor, Malaysia, for the years 2009 (training) and 2010 (testing). The second dataset covered daily consumption in Germany from 2012–2015 (training) and 2016–2017 (testing). These datasets [157], differing in temporal granularity (from hourly to yearly temporal samples), allowed the researchers to assess the models ability to generalise across distinct types of forecasting problems.

The results over-performed the state-of-the-art on the STLF task. On the Malaysian dataset, PLCNet achieved an exceptional Accuracy (AC) of 98.23%, while on the German dataset, Accuracy improved from 83.17% to 91.18% compared to previous models. Furthermore, the parallel design showed faster training times compared to deeply stacked sequential architectures. It was also more robust when forecasting over different time steps, such as predicting the next 24 or 48 hours, or even up to 10 days ahead. Also, the model handled both stationary and non-stationary data well, reducing the need for pre-processing steps common in traditional time-series models [135].

Research illustrates that a parallel DL architecture, where temporal and spatial patterns are learned simultaneously rather than sequentially, can offer significant gains in Accuracy, adaptability, and efficiency [158].

In this master thesis, we introduce the Parallel Network CNN - LSTM (ParaNet-CNN-LSTM) architecture for the first time in the context of environmental acoustic soundscapes, following a design concept similar to the PLCNet model [135]. Here, we present the architecture of our ParaNet-CNN-LSTM.

ParaNet-CNN-LSTM Architecture

The Parallel Network CNN - LSTM (ParaNet-CNN-LSTM) model integrates the complementary strengths of both CNN and LSTM networks by processing the same input in parallel through two independent branches. The dual-pathway structure enables the model to capture both local feature representations and long-range temporal dependencies inherent in acoustic data.

The CNN branch specialises in identifying spatially invariant patterns within the input, while the LSTM branch is dedicated to modelling sequential dynamics over time. After

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

the independent feature extraction, the outputs of both branches are concatenated and passed through a Fully Connected layer, which performs the final classification task.

CNN Branch

The CNN branch receives as input a three-dimensional tensor of shape (N, T, F), where N denotes the batch size, T corresponds to the number of time steps in the sequence, and F represents the number of input features per time step (`input_size`). This input is obtained by permuting the original input tensor to match the expected format of 1D convolutional layers, which require the shape (N, C, L), where C is the number of channels and L the sequence length. In this context, F becomes the number of channels, and T the sequence length.

The convolutional pathway applies a sequence of operations designed to extract localised, hierarchical features from the input signal across the temporal dimension. This CNN module is identical as the CNN module of the SeqNet-LSTM-CNN model (See Section 3.2.2).

However, in this parallel configuration, the last linear transformation projects the output into a latent space of size 128, differing from the SeqNet-LSTM-CNN model. Consequently, the CNN branch serves as a local pattern encoder and outputs a feature vector of shape $(N, 128)$, matching the dimensionality of the parallel LSTM branch. This ensures that both representations contribute equally to the final classification task when they are concatenated.

LSTM Branch

The LSTM branch is designed to capture long-term temporal dependencies in the sequential data. It processes the original input tensor in its native shape (N, T, F), where N represents the batch size, T is the number of time steps (sequence length), and F (`input_size`) is the number of features per time step. This input is fed into a stacked LSTM network consisting of two layers (`num_layers` $= 2$) with a hidden state dimensionality of 128 (`hidden_size`). The stacked LSTM architecture enables the model to retain and update internal memory across time, effectively modelling sequential dependencies and patterns that evolve over multiple time steps.

The LSTM outputs the hidden state for each time step in the sequence, from which the hidden state corresponding to the final time step, serving as a condensed representation of the entire input sequence. This vector has a shape of $(N, 128)$ and is subsequently passed through a FC linear transformation layer projecting it into a 128-dimensional space, the approach is identical as the one used in all the proposed models (See Sections 3.2.1, 3.2.2).

The final output aligns in shape with the representation produced by the parallel CNN branch, enabling their concatenation prior to the final classification stage. This parallel configuration allows the model to leverage both temporally-aware and spatially-localised feature representations for robust sequence classification.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Concatenation and Output

The ParaNet-CNN-LSTM model orchestrates the simultaneous processing of the input sequence through both the convolutional (CNN) and recurrent (LSTM) branches, enabling the fusion of local spatial features and long-term temporal dependencies.

The input tensor, originally of shape (N, T, F) is split and formatted appropriately for each processing branch. The CNN branch requires the input to be in the format (N, F, T), so the tensor is first permuted before being passed through the convolutional feature extractor. This produces a fixed-size representation per instance of shape $(N, 128)$.

In parallel, the same input (in its original (N, T, F) format) is passed through the LSTM, which returns the hidden states across all time steps. Only the hidden state at the final time step is retained, resulting in a tensor of shape $(N, 128)$, which is then projected through a fully connected layer to match the output dimensions of the CNN branch.

These outputs from both branches, each of shape $(N, 128)$, are concatenated along the feature axis, forming a combined representation of shape $(N, 256)$ [159]. This joint feature vector is finally passed through a Fully Connected linear layer that maps it to the output space of size equal to the number of target classes (`num_classes`), producing logits of shape $(N, \texttt{num\_classes})$ suitable for classification.

Concatenation: The outputs from the CNN and LSTM branches (both 128-dimensional) are concatenated along the feature dimension, resulting in a 256-dimensional vector.

FC Linear Layer: This operation maps the 256 combined features to `num_classes` output units, producing the final logits for classification.

Output: Logits of shape $(N, \text{num\_classes})$

At the end of this parallel workflow, the final FC layer performs the classification by projecting the 256-dimensional feature vector, composed by concatenating two 128-dimensional outputs from each branch, into a vector of class logits, where each logit corresponds to a specific target category within the defined Regions (Reference Forest, Pasture, Natural Regeneration, and Plantation).

3.2.4 Reference Models for Benchmarking

An essential aspect of this master thesis research is the comparative evaluation of the proposed models against established architectures. This step is crucial to ensure a rigorous scientific approach and to identify the strengths and limitations of our FOREST, and proposed model architectures. For this purpose, we selected the following baseline models as references for benchmarking:

- Simple Support Vector Machine (SVM) [156]

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

- Simple Convolutional Neural Network (CNN) [150]
- Simple Long Short-Term Memory (LSTM) [153]
- ResNet34 [160], [161]

With the selected models, we aim to explore how the dataset responds to different types of architectures, including a linear inference model (SVM) and the individual components of our proposed architecture (i.e., CNN and LSTM). Additionally, we compare our approach with a more complex and well-established architecture for audio classification and image recognition, namely the ResNet34 model [162]–[165]. All benchmark results will be analysed and discussed further in Section 3.4, and Section 4.2.

3.3 Model Training Procedure

The training procedure for all the proposed and reference models is the same, and follows a standard supervised learning setup in PyTorch [44], involving forward and backward [166] passes, loss computation, and parameter optimisation over multiple epochs [167]. The model is trained using a multi-class classification objective with cross-entropy loss [168] and optimised with the Adam optimizer [169]. During training, the dataset is iteratively fed into the model in mini-batches using a DataLoader [170], and the model's parameters are updated through backpropagation [166].

Several training durations were explored, ranging from 3 to 91 epochs. A training duration of 11 epochs was found to provide a suitable balance between model performance and computational efficiency [164]. The training process is designed to record each model's performance in JSON-formatted dictionaries, enabling serialisation within the HPC filesystem for subsequent analysis.

3.3.1 Training and Testing Data Split

The dataset was partitioned into training and testing sets using an 80/20 split, ensuring that 80% of the data was used for training and 20% for testing [171]. To address the issue of class imbalance within the dataset, stratified sampling was employed during this partitioning process [172]. The stratification strategy preserved the original distribution of the target classes in both the training and testing subsets, which is crucial for obtaining reliable and unbiased evaluation metrics.

Prior to model training, the input features were normalised independently using a min-max scaling method adapted for time series data [173]. This normalisation scales each feature to a fixed range (between 0 and 1) enhancing model convergence and stability.

Finally, the normalised data and corresponding labels were converted into PyTorch tensors and organised into `TensorDataset` [174] objects to facilitate efficient loading and batching during the training and evaluation phases, using the advantages of the PyTorch library [44].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

3.3.2 Training Configuration

For this research, we utilised the Computational Power and Graphics Processing Unit (GPU) resources of two High-Performance Computing (HPC) Systems: Izar [71], operated by Scientific IT & Application Support (SCITAS) at the École Polytechnique Fédérale de Lausanne (EPFL), and VSC-5 [72], provided by the Austrian Scientific Computing (ASC) at the Technische Universität Wien (TU Wien).

The training process follows a standard supervised learning workflow. For each model, the training loop iterates over the dataset for a predefined number of epochs. In each epoch, batches of input data and corresponding labels are retrieved from the `DataLoader` [170]. These inputs are transferred for computation GPU and passed through the model to generate predictions. The loss between predictions and ground truth labels is computed using a cross-entropy loss function [168]. Gradients are then calculated via backpropagation [166], [167], and the model parameters are updated using the Adam optimizer [169]. This cycle continues until all epochs are completed, gradually minimizing the loss and improving the model's performance on the training data.

Loss Function (Training Criteria)

The cross-entropy loss is the criterion for the proposed loss function used for multi-class classification tasks with C classes. It expects the input to be unnormalised logits. The target labels should be class indices in the range $[0, C]$ [168].

The unreduced loss for class indices (label form) is defined as:

$$\begin{aligned} \ell(x, y) &= L = \{l_1, \ldots, l_N\}^\top, \\ l_n &= -w_{y_n} \log \left(\frac{\exp(x_{n,y_n})}{\sum_{c=1}^{C} \exp(x_{n,c})} \right) \cdot \mathbf{1}\{y_n \neq \texttt{ignore\_index}\}. \end{aligned} \tag{3.9}$$

Unreduced weighted cross-entropy loss function with ignored indices [168].

If the reduction method is set to `'mean'` (default) or `'sum'`, the final reduced loss is:

$$\ell(x, y) = \begin{cases} \dfrac{1}{\sum_{n=1}^{N} w_{y_n} \cdot \mathbf{1}\{y_n \neq \texttt{ignore\_index}\}} \displaystyle\sum_{n=1}^{N} l_n, & \text{if reduction} = \texttt{'mean'} \\ \displaystyle\sum_{n=1}^{N} l_n, & \text{if reduction} = \texttt{'sum'} \end{cases} \tag{3.10}$$

Reduced cross-entropy loss function implementation [168].

We implemented the PyTorch cross-entropy loss function with default parameters [168].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Optimiser Algorithm for Loss Function

An Adaptive Moment Estimation (Adam) optimiser is implemented with a learning rate of 0.001, the cross-entropy loss as the input objective-function, and all other default parameters. This optimiser adaptively adjusts learning rates for each parameter, promoting faster convergence as follow [169], [175].

Algorithm 3.1: Adam Optimisation Algorithm [169], [175]

Input:
Learning rate γ,
exponential decay rates = β_1, β_2,
initial parameters = θ_0,
objective function = $f(\theta)$,
weight decay = λ,
flags = `amsgrad`, `maximize`,
epsilon = ϵ

Output: Updated parameters θ_t

Initialise: $m_0 \leftarrow 0$ (first moment), $v_0 \leftarrow 0$ (second moment), $\hat{v}_0^{\max} \leftarrow 0$

for $t = 1, 2, \ldots$ **do**
  if *maximize* **then**
    $g_t \leftarrow -\nabla_\theta f_t(\theta_{t-1})$;
  else
    $g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$;
  end
  if $\lambda \neq 0$ **then**
    $g_t \leftarrow g_t + \lambda\theta_{t-1}$; // Weight decay
  end
  $m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t$;
  $v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2) g_t^2$;
  $\hat{m}_t \leftarrow \frac{m_t}{1-\beta_1^t}$;
  if *amsgrad* **then**
    $\hat{v}_t^{\max} \leftarrow \max(\hat{v}_{t-1}^{\max}, v_t)$;
    $\hat{v}_t \leftarrow \frac{\hat{v}_t^{\max}}{1-\beta_2^t}$;
  end
  else
    $\hat{v}_t \leftarrow \frac{v_t}{1-\beta_2^t}$;
  end
  $\theta_t \leftarrow \theta_{t-1} - \gamma \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$;
end
return θ_t;

Adam is an adaptive learning rate optimisation algorithm designed specifically for

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

training deep NNs. This algorithm is an efficient first-order optimiser that adaptively adjusts learning rates for each parameter based on estimates of the first and second moments of the gradients [169]. Also, it is well-suited for training DL models like CNNs and LSTMs, in particular with noisy or non-stationary input data [175].

Number of Epochs

An epoch is a complete pass through all of the training data. For this research, we experimented with a range of epochs from 3 to 91. Our empirical results indicated that 11 epochs strike the best balance between learning efficiency and energy consumption, so we avoid unnecessarily burning trees from the FOREST for this master thesis research [176]. This value was used to evaluate each model architecture. During each epoch, the entire training dataset is processed once. Multiple epochs enable the progressive refinement of model weights through repeated exposure to the complete training data.

Mini-batch Iteration

Each mini-batch `(x, y)` is loaded from the `train_loader`, where x is the input tensor of shape (N, T, F) and y is the corresponding label tensor. These tensors are moved to the appropriate computation device (GPU).

Gradient Reset

The existing gradients need to be cleared with `optimizer.zero_grad()` before computing new gradients. This clearance-operation is necessary because PyTorch accumulates gradients by default. Using this function we aim to lower memory footprint, and improve the prediction performance [177].

Forward Pass

The model produces predictions using `y_pred = model(x)`. This involves processing the input through all the layers of both the CNN & LSTM models (as previously described in Section 3.2) [167]. As a result, we obtain logits vector of shape (N, C). where N represents the batch size, and C the vector with the prediction rate for each class.

Loss Computation

The loss between the predicted logits and the true labels is calculated using `loss = criterion(y_pred, y)`, which for our research is cross-entropy [168]. This function computes value-measures of the distances of the model's predictions with respect the ground truth [167].

Backward Pass

The gradients of the loss adjust with respect to all model parameters are computed using `loss.backward()` [166]. This function initiates backpropagation through the model graph.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Parameter Update

The model's parameters are updated using `optimizer.step()` [178], which applies the computed gradients via the Adam optimisation algorithm [169].

Model Serialisation

After many hours of computation, it is time to harvest the fruits of the FOREST. Accordingly, the model weights that achieved the best performance are saved using `torch.save()` for future experiments [179].

3.4 Model Evaluation

In this Section, we describe the evaluation metrics (see Section 3.4.1), and expose the preliminary results of the empirical experiments (see Section 3.5) conducted to evaluate the overall performance of the models introduced for this master thesis. We employed the curated dataset from Section 2.3.2, obtained after the preprocessing workflow from Section 2.3.1, to train each model and ensure a scientifically valid comparison, and evaluation.

As benchmarks, we trained three baseline models: a Support Vector Machine (SVM), a Simple Convolutional Neural Network (CNN), and a Simple Long Short-Term Memory (LSTM) network. Additionally, we implemented a customised 1Dimensional (1D) ResNet34 [160] architecture tailored for the audio classification task. These four (4) baseline models serve as references against which we compare our proposed architectures: Sequential Network CNN - LSTM (SeqNet-CNN-LSTM), Sequential Network LSTM - CNN (SeqNet-LSTM-CNN), and Parallel Network CNN - LSTM (ParaNet-CNN-LSTM).

3.4.1 Evaluation Metrics

We evaluated the performance of each model by first setting the PyTorch model objects to evaluation mode, and then executing their prediction mechanism using the `model.forward()` function [167], followed by the corresponding evaluation procedures. This was performed using the previously split `test_loader` dataset (see Section 3.3.1), which was passed through each model to generate the `y_pred` predictions. These predictions are then used to compute the confusion matrix and the Accuracy (AC) metric, both of which are subsequently serialised into JSON-formatted dictionaries. The overall process follows the three (3) steps outlined below.

1. **Activate Evaluation Mode**

   The evaluation loop is wrapped within `torch.no_grad()` to reduce memory usage and accelerate computations during inference [177]. Then, the model is set to evaluation mode via `model.eval()`, which disables training-specific layers such as Dropout and Batch Normalisation that is functionally equivalent to set `model.train(False)` [167].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

2. **Predict and Calculate Accuracy**

   The classification correctness for each class is determined using the following metrics.

   True Positives (TP): Correctly predicted instances of the class.

   True Negatives (TN): Correctly predicted instances not belonging to the class.

   False Positives (FP): Incorrectly predicted instances as belonging to the class.

   False Negatives (FN): Instances of the class incorrectly predicted as another class.

   The model Accuracy is calculated as follows.

   1. Each mini-batch (x, y) from the test set is transferred to the GPU.
   2. The predictions are obtained using `y_pred = model(x)` [167].
   3. The predicted class labels are derived using `torch.max` [180].
   4. The correct predictions are Aggregated to calculate AC using the Equation 3.11.

$$\text{Accuracy (AC)} = \frac{\text{Number of Correct Predictions (TP + TN)}}{\text{Total Number of Samples}} \times 100 \tag{3.11}$$

Accuracy (AC) calculation equation.

3. **Aggregate Result**

   For each model, the following metrics are computed and serialised.

   `EPOCHS`: Number of training epochs.

   `AC`: Classification Accuracy (in percentage).

   `TP`, `TN`, `FP`, `FN`: Confusion matrix components.

These metrics are serialised into JSON-formatted dictionaries for further analysis and result reporting. The customised implementation of the evaluation routine is summarised in the `test()` function, which processes a list of models and the test dataset, returning performance statistics per model.

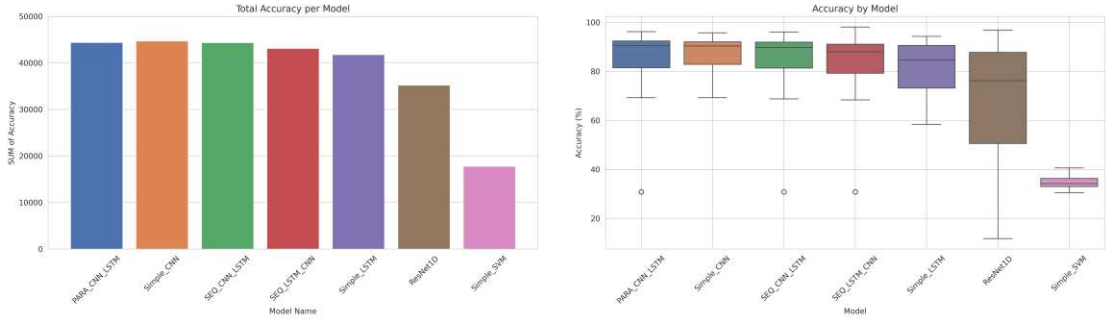
3.5 Preliminary Results

In order to asses model performance in the task of environmental soundscape classification, we analyse the results of systematically training seven (7) models using different feature groupings from the curated dataset presented in Section 2.3.2. Each configuration includes a fixed subset of seven (7) features and a variable subset of nine (9) features, combined in

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

all possible non-empty subsets, yielding 511 unique combinations. Each combination is appended to the fixed part, resulting in feature sets of size $7 + r$, where $r \in \{1, 2, \ldots, 9\}$, as described in the experimental setup in Section 4.1. In total, 3577 training runs are executed, and the performance of each run is recorded in JSON-formatted dictionaries containing the evaluation metrics and associated metadata, as detailed in Section 3.4.1.

Here, we present the preliminary Aggregated (AGG) results summarising the performance of all models introduced in Section 3.2. The aim is to provide an initial overview of how well each architecture performs on the environmental soundscape classification task.



(a) Aggregated (AGG) Accuracy (AC) by Model (b) Accuracy (AC) Distribution by Model

Figure 3.1: Model Performance Overview

The Figure 3.1a presents a comparative analysis of the total Aggregated (AGG) Accuracy (AC) achieved by all evaluated models: ParaNet-CNN-LSTM, Simple-CNN, SeqNet-CNN-LSTM, SeqNet-LSTM-CNN, Simple-LSTM, ResNet1D, and Simple-SVM. The overall Aggregated (AGG) AC reflects the cumulative performance over all 511 training and testing runs per model, where a perfect performance in each run would yield a MAX score of 51100. One can clearly differentiate models with higher total AC, such as ParaNet-CNN-LSTM, Simple-CNN, SeqNet-CNN-LSTM, and SeqNet-LSTM-CNN, from those with comparatively lower performance, including Simple-LSTM, ResNet1D, and Simple-SVM. This distinction suggests that models incorporating convolutional layers generally outperform those without, particularly in the context of this particular task of environmental soundscape classification.

The models within the CNN-LSTM family, namely ParaNet-CNN-LSTM, SeqNet-CNN-LSTM, and SeqNet-LSTM-CNN, demonstrate consistently strong performance due to their ability to jointly capture spatial and temporal characteristics of the input data. Among simpler architectures, Simple-CNN outperforms both Simple-LSTM and the CNN-LSTM variants in terms of cumulative AC, indicating that even basic convolutional models can be highly effective for this task.

In contrast, Simple-SVM exhibits the lowest total AC, underscoring the limitations of traditional machine learning methods such as SVM in modelling complex, high-

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

dimensional acoustic patterns. Its linear decision boundaries appear insufficient to capture the variability inherent in environmental audio data.

Although ResNet1D is a Deep Learning (DL)-based model, its performance is relatively moderate. This may be attributed to its architectural origin in image classification using 2-Dimensional (2D) convolutions, which may not optimally generalise to 1Dimensional (1D) audio signals.

Overall, the best-performing models are ParaNet-CNN-LSTM, SeqNet-CNN-LSTM, and Simple-CNN which highlight the advantages of convolutional operations, either alone or in combination with recurrent layers. The noticeable variation in performance among the CNN-LSTM variants also suggests that specific architectural design choices, such as the order or integration of layers, play a significant role in optimising model effectiveness for soundscape classification.

The Figure 3.1b presents the Accuracy (AC) distribution across all models over the 511 training and testing runs. Among these, the models exhibiting the most similar performance distributions are ParaNet-CNN-LSTM, Simple-CNN, SeqNet-CNN-LSTM, and SeqNet-LSTM-CNN. Notably, three of these models integrate both convolutional and recurrent layers, and all of them include convolutional layers arranged sequentially, differing only in the order of CNN and LSTM components.

In contrast, the most distinct Accuracy distributions are observed in Simple-LSTM, ResNet1D and SVM. These models represent fundamentally different modelling paradigms. Any of these models employ an explicit convolutional layer. The ResNet1D model employs deep residual learning without explicit temporal modelling, while SVM is a linear Machine Learning (ML) model with no deep feature extraction capabilities.

The box-plots in Figure 3.1b indicate that models with convolutional layers, such as ParaNet-CNN-LSTM, Simple-CNN, SeqNet-CNN-LSTM, and SeqNet-LSTM-CNN, achieve consistently high performance, with interquartile ranges (Q1 to Q3) approximately spanning the 80–90% Accuracy range. This consistency suggests that these architectures effectively capture spatial features regardless of whether temporal modelling occurs before or after spatial processing. The Simple-LSTM model, which relies solely on temporal information, achieves modest performance with a broader interquartile range, roughly between 70% and 90%.

These insights suggest that while the sequence of convolutional and recurrent layers may influence performance slightly, both layers contribute substantially and complementarily to the task. The architectural flexibility in sequencing does not lead to drastic changes in accuracy, though models starting with spatial extraction, such as ParaNet-CNN-LSTMm Simple-CNN, and SeqNet-CNN-LSTM, show marginal improvements over all the configurations.

The performance of the parallel architecture, ParaNet-CNN-LSTM, slightly surpasses its sequential counterparts, suggesting that parallel extraction of spatial and temporal features results in richer and less constrained feature representations. In contrast to

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

sequential models, where one module transforms the output of the other, the parallel architecture could avoid bottlenecks and preserve distinct spatial and temporal information simultaneously.

The ResNet1D model achieves a median Accuracy of around 80%, benefiting from deep residual connections for complex feature extraction. However, the absence of temporal modelling likely limits its performance when compared to the hybrid CNN-LSTM family models. While ResNet1D outperforms traditional models, it is surpassed by all DL models that integrate both spatial at least a simple convolutional layer.

In addition, the ResNet1D model shows very high variance in its Accuracy distribution. Although it reaches one of the highest maximum accuracies, its performance is unstable, with a wide spread between the lower and upper quartiles. Its minimum Accuracy drops below 20%, while its maximum approaches 100%. The median Accuracy remains below 80%, and the interquartile range spans from approximately 50% to 70%. This variability may stem from sensitivity to data splits, overfitting, insufficient data, or unstable training dynamics.

Furthermore, traditional approaches like SVM exhibited significantly lower performance, highlighting the limitations of shallow models on structured and high-dimensional data. These findings suggest that DL-based architectures integrating both convolutional and recurrent components, particularly in parallel configurations, should be prioritised for High-Performance Computing (HPC) applications in environmental soundscape analysis. The ability of such models to learn complementary representations enables them to generalise more effectively and deliver consistent results, even in the presence of complex acoustic patterns.

Among all models, four primary insights emerge. First, combining CNN and LSTM components is crucial. Models that integrate both spatial and temporal processing outperform those that rely on only one type of feature extraction, underscoring the need for hybrid architectures in soundscape classification. Second, the ParaNet-CNN-LSTM model consistently outperforms all others, including the sequential variants, likely due to richer multi-path feature representation and reduced information loss of the parallel integration of convolutional and recurrent layers yielding the best results. Third, the ordering in sequential architectures appears to have a limited effect on performance. Both SeqNet-CNN-LSTM and SeqNet-LSTM-CNN perform similarly, indicating that the model is robust to the order in which features are extracted, although extracting spatial features first yields slightly better results. Fourth, traditional ML-based approaches are significantly outperformed. The poor Accuracy of the SVM model illustrates the limitations of shallow models when dealing with complex spatio-temporal patterns.

In summary, both charts in Figure 3.1 highlight the effectiveness of hybrid DL-based models, particularly those with parallel CNN and LSTM pathways, in maximising classification performance for environmental soundscape tasks. These model architectures that are capable of simultaneously capturing spatial and temporal dependencies offer a clear advantage in soundscape classification tasks. The ParaNet-CNN-LSTM model

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

consistently achieved the highest levels of accuracy, demonstrating superior robustness and feature extraction capacity across diverse test scenarios. Sequential hybrid models, such as SeqNet-CNN-LSTM and SeqNet-LSTM-CNN, also performed well, although with slightly lower accuracy, making them suitable alternatives under computational or architectural constraints.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 4

Evaluation

Chapter 4 aims to evaluate the performance of the proposed model architectures introduced in Section 3.2, as well as the influence of the each Ecological Acoustic Indices (EAIs) described in Section 2.2 on the classification performance. The primary objective is to identify the most suitable model architecture and the optimal combination of EAI features for the given task. To this end, we present the experimental setup in Section 4.1, detailing the empirical procedures used to assess both the model performance and the individual contributions of each EAI within the curated dataset from Section 2.3. Subsequently, Section 4.2 reports the outcomes of these experiments, highlighting the most appropriated model architecture, the specific contribution of each EAI on the models' Accuracy, and the most appropriated set of EAIs for the Soundscape classification task.

4.1 Experiments Setup

In this section, we describe the experimental setup designed to evaluate the performance of the proposed models that aims to identify the most suitable model architecture and to assess the influence of each specific Ecological Acoustic Indice (EAI) on model performance in the context of soundscape classification.

To achieve this, we conduct a series of empirical evaluations using the curated dataset introduced in Section 2.3.2 and its corresponding EAI. Each model presented in Section 3.2 was tested with a varying number of input features, ranging from 8 to 16, in order to explore the impact of different feature combinations on classification performance.

The first feature is the target variable (also known as golden label). Then, the following three (3) features represent metadata associated with each soundscape. The subsequent four (4) features are statistical descriptors, and the final eight (8) features correspond to Ecological Acoustic Indices.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

1. REGION
2. SOUNDSCAPE_ID
3. BAND_ID
4. SECOND
5. MEAN
6. MEDIAN
7. Summation (SUM)
8. Maximum (MAX)
9. Acoustic Complexity Index (ACI)
10. Bioacoustic Index (BET)
11. Temporal Median (MMM)
12. Number of Peaks (NPP)
13. Frequency Entropy (HFQ)
14. Temporal Entropy (HTP)
15. Combined Entropy (HHH)
16. Acoustic Evenness Index (AEI)

Following these experiments, we analyse the results to determine the most suitable model architecture and to evaluate how the Number of EAIs influences model performance in the specific task of environmental soundscape classification. Additionally, we aim to identify an appropriate Number of EAIs that optimises the overall feature set.

We systematically explore different feature groupings for model training by generating all possible combinations of a variable subset of features. From the total pool of 16 features, a fixed subset (`fixed_part`), consisting of the first seven (7) entries in the curated dataset from Section 2.3.2, is kept constant. The remaining nine (9) features (`variable_part`) are combined in all possible non-empty subsets, resulting in $2^9 - 1 = 511$ unique combinations. Each combination is then appended to the `fixed_part`, yielding feature sets of size $7 + r$, with $r \in \{1, 2, \ldots, 9\}$, as summarised in Table 4.1.

Variable r Features (EAIs)	Total Features	Total Unique Combinations ($\binom{9}{r}$)
1	8	9
2	9	36
3	10	84
4	11	126
5	12	126
6	13	84
7	14	36
8	15	9
9	16	1
Total	16	**511**

Table 4.1: Number of Feature Combinations for Each Value of r (EAIs)

Consequently, each model is trained and evaluated 511 times, once for every feature combination, leading to a total of $7 \times 511 = 3577$ runs across the seven (7) models. Each run outputs a JSON-formatted dictionary containing the performance metrics and associated metadata, as described in Section 3.4.1.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

4.2 Experiments Result

As presented in Section 3.5, the Aggregated (AGG) Accuracy (AC) results provided a general overview of model performance across all architectures, highlighting how architectural choices influence soundscape classification. Building on this, we further extended the analysis to assess the impact of varying the number of EAIs within the dataset across all model architectures.

Next, we narrow the analysis to a subset of models that demonstrate consistently stronger performance. The Support Vector Machine (SVM) model is excluded, as its linear prediction mechanism cannot adequately capture the statistical distribution of features in this dataset. Similarly, the ResNet1D model is omitted due to its high variability across different numbers of EAIs, which introduces noise into the evaluation. Excluding these models ensures a more stable and representative assessment of how varying the number of EAIs influences the performance of CNN-based architectures.

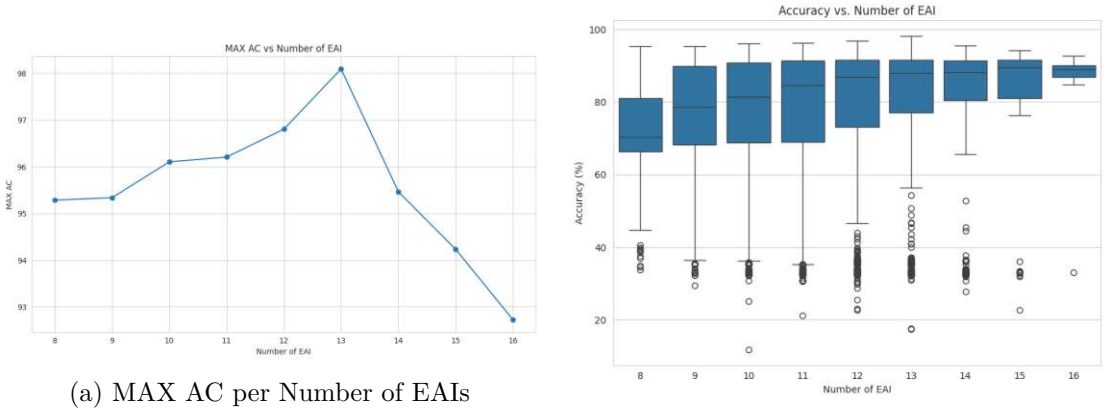
Furthermore, we evaluate and compare the results across the CNN-based models, presenting preliminary findings on how the number of EAIs influences the performance of both the reference models and the proposed architectures.

Finally, we perform a focused subsample analysis by narrowing the number of EAIs, first considering the range from 10 to 16, and then further restricting it to the range from 10 to 13. These subsamples allow us to concentrate on identifying the most suitable and effective DL-based model architecture for the task of ecological soundscape classification.

4.2.1 Impact of the Number of Ecological Acoustic Indices

For this experiment, we evaluate all 511 feature combinations across all models, resulting in a total of 3577 runs. We analyse the distribution of AC across the total number of features ranging from 8 to 16, in order to identify performance peaks aggregated over all model architectures without exclusion. First, we plot the maximum Accuracy achieved for each feature count in Figure 4.1a. Then, we visualise the distribution of Accuracy using boxplots grouped by the total number of features, as illustrated in Figure 4.1b.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN



(a) MAX AC per Number of EAIs

(b) AGG AC Distribution per Number of EAI

Figure 4.1: Accuracy (AC) Overview per Number of EAIs

The Figure 4.1a illustrates the Maximum (MAX) AC obtained for each total number of features, ranging from 8 to 16. The Accuracy steadily increases from 8 to 11 features, followed by a more pronounced improvement peaking at 13 features. This peak suggests an optimal feature set size, beyond which additional features do not contribute further gains and may introduce redundancy or overfitting effects. Notably, Accuracy slightly declines beyond 13 features. This trend reflects diminishing returns in performance and emphasises the existence of an optimal performance around 11–13 features, where feature interactions likely provide maximal synergy.

However, beyond 13 features the trend begins to plateau, and in some cases, a slight decline is observed. This suggests diminishing returns when more features are added, likely due to increased redundancy, or possible noise. The reduced diversity in feature sets at higher counts could lead to less varied learning scenarios, ultimately stabilising or even degrading performance slightly. One can observed a critical balance while incorporating a higher number of features can improve Accuracy, there exists an optimal range, approximately 11 to 13 total number of features, that provides the best trade-off between model complexity and predictive performance.

The Figure 4.1b provides a complementary view by showing the full distribution of AC across all 3577 model evaluations, grouped by total feature count. The boxplots illustrate how both the median and the spread of Accuracy values change as more features are included. At lower feature counts (8 to 11), the Interquartile Range (IQR) is noticeably wider, indicating a higher variability in model performance. This suggests that, with fewer features, the models are more sensitive to the specific combination used, leading to greater inconsistency across runs.

As the number of features increases, particularly from 12 to 16, the IQR narrows noticeably, and the overall distribution becomes more compact. The median Accuracy approaches the upper performance boundary, suggesting enhanced consistency and robustness. Also, The

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

frequency and severity of outliers decrease with higher feature counts, though occasional low-performing combinations persist even at 16 features. This trend reflects the benefit of richer feature combinations, but it is also partially influenced by the decreasing number of available combinations as feature count increases. With fewer data points at higher feature counts, the compactness of the IQR may not solely reflect improved model performance but also reduced variability in the sample space.

Despite this, the upper quartile remains close to the maximum observed AC, and the shrinking lower tails indicate fewer low-performing outliers. These patterns suggest that larger feature sets tend to yield more stable and reliable results. However, occasional low-performing runs persist even at higher feature counts, likely due to residual variability in the specific feature subsets selected. Together, these findings underscore a balance between feature richness and combination diversity in driving both AC and consistency.

The Figure 4.1 shows that the overall models' performance generally improves with more features, up to an optimal range between 11 and 13. Within this range, the models reach their highest AC with low variability. Beyond this point, gains become marginal, and the risk of overfitting increases. One can highlighted the importance of selecting a balanced set of EAIs, as adding too many features may reduce diversity and limit overall performance.

4.2.2 Impact of the Number of EAIs per Model Architecture

Building upon the preliminary results from Section 3.5 and the analysis of the number of EAIs on the Aggregated model performance in Section 4.2.1. In this Section, we analyse the specific impact of the number of EAIs on each model architecture individually. The aim is to understand how the distribution of the Accuracy varies per model as the number of EAIs increases from 8 to 16, using the same experimental setup described in Section 4.1. The resulting performance distributions are shown in Figure 4.2.

The Figure 4.2 shows the statistical distribution of the Accuracy (AC) among each individual model architecture as a function of the number of EAIs from 8 to 16. The boxplots illustrate central tendencies, variability, and robustness of each model's performance. The evaluated models include ParaNet-CNN-LSTM, Simple-CNN, SeqNet-CNN-LSTM, SeqNet-LSTM-CNN, Simple-LSTM, ResNet1D, and Simple-SVM.

Most of the models display an increase in median Accuracy with a higher number of EAIs, suggesting improved performance as more features are incorporated in the dataset. This trend is more pronounced and consistent in hybrid and CNN-based architectures, where the Interquartile Range (IQR) tends to narrow around 11 to 13 number of features, indicating enhanced stability. In contrast, Simple-Non-CNN-based models exhibit wider Accuracy distributions and more outliers, reflecting sensitivity to feature subset variations and less robust learning.

The outliers are mainly present in lower Accuracy values. These are common across all the models but diminish in frequency, and magnitude as the number of EAIs increases,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

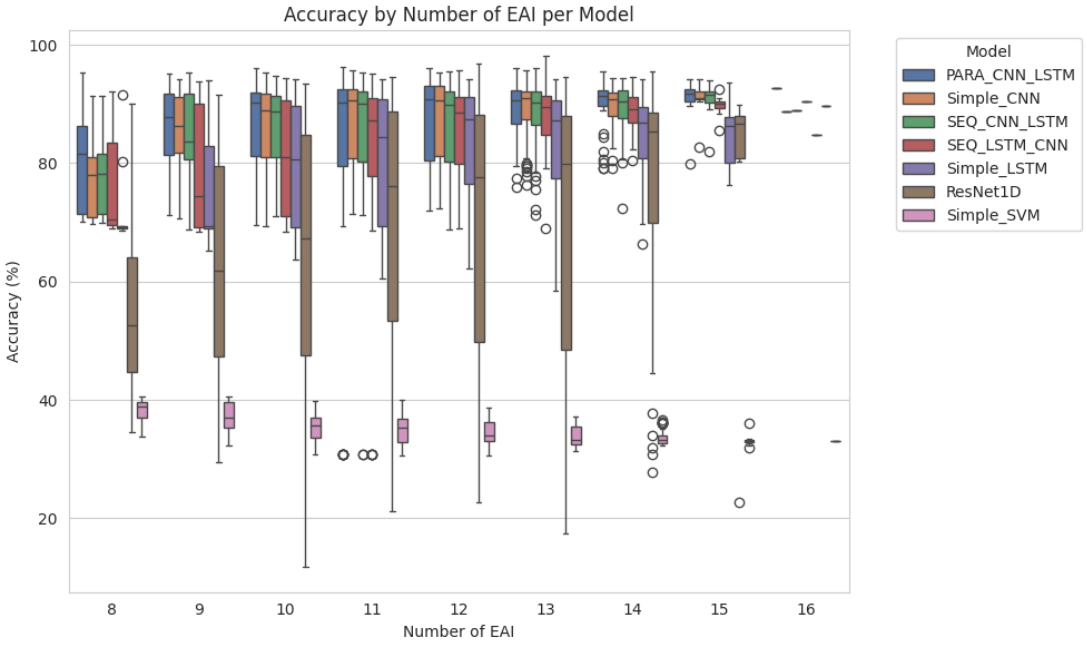


Figure 4.2: Accuracy (AC) Distribution per Model Architectures by Number of EAIs

particularly in architectures capable of effectively leveraging spatial, and temporal patterns from the audio signals.

In particular, the ParaNet-CNN-LSTM achieves consistently high median Accuracy with a steady upward trend and moderate variability across the range of 8 to 16 number of EAIs, along with relatively few low outliers. This indicates stable and reliable performance. Similarly, the Simple-CNN, SeqNet-CNN-LSTM, and SeqNet-LSTM-CNN models display clear gains in median Accuracy and reduced spread, characterised by compact IQR distributions and minimal outliers. These patterns reflect robust and consistent performance across varying feature set sizes.

In contrast, model architectures such as ResNet1D, SeqNet-LSTM-CNN, and Simple-LSTM exhibit higher variability, with wider IQRs across all range-size of EAIs, indicating inconsistent Accuracy gains despite the increase in feature count. Among these, the SeqNet-LSTM-CNN shows the highest AC performance among all the models with 13 number of EAIs, and moderate improvements in median Accuracy, particularly from 11 EAIs onward. However, it also displays greater variability and more frequent outliers compared to the other models from the CNN-based architectures, suggesting less robust behaviour. Potentially, due to the order of the learning process which extract the temporal patterns first by the LSTM layer, and then these are re-learned as for spatial patterns by the CNN layers.

Additionally, ResNet1D maintains relatively wide IQRs across the entire range of EAIs, reflecting persistent variation in Accuracy with low performance. Simple-SVM records

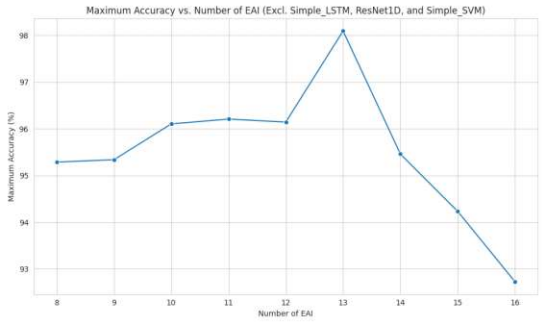
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

the lowest Accuracy values among all models, with performance deteriorating as the number of EAIs increases. This highlights their unstable behaviour and limited suitability for the soundscape classification task, as it was detailed in Section 3.5.

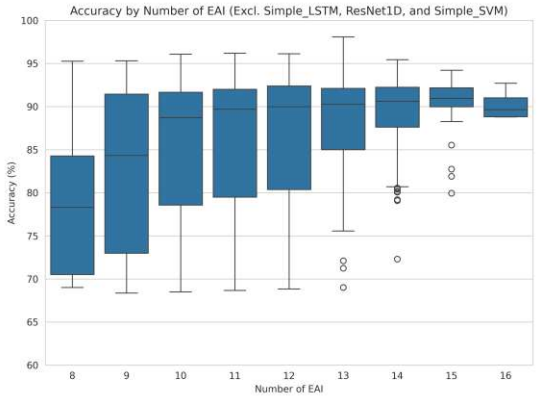
This comparison highlights that model architecture plays a significant role in shaping both the AC and its stability as the number of EAIs increases, consistent with the observations reported in Section 3.5. The CNN-based architectures, particularly ParaNet-CNN-LSTM, SeqNet-CNN-LSTM, and Simple-CNN, consistently demonstrate superior performance in terms of higher median, more compact AC distributions, and fewer extreme outliers. Although, increasing the number of EAIs generally enhances performance, the degree of improvement reaches a peak at the number of 13 EAIs. These findings underscore the importance of selecting an appropriate combination of informative features and robust model architectures to ensure reliable and accurate classification outcomes.

4.2.3 Selection of the Most Appropriated Model Architecture

Consequently, based on the insights discussed above in Section 4.2.2, we exclude model architectures that consistently exhibit lower Accuracy (AC) performance and reduced stability, namely ResNet1D, Simple-LSTM, and Simple-SVM. By narrowing the scope to the remaining better-performing models, we aim to more clearly identify the most appropriate number of EAIs features. Also, this focused approach allows for a more precise and consistent evaluation, thereby paving the path for the third subsequent experiment presented in Section 4.2.4.



(a) MAX AC per Number of EAIs



(b) AGG AC Distribution per Number of EAIs

Figure 4.3: Accuracy (AC) as a Function of the Number of EAIs Excluding Simple-LSTM, Simple-SVM, and ResNet1D

The Figure 4.3a presents the MAX Accuracy reached among all the Selected Models (ParaNet-CNN-LSTM, Simple-CNN, SeqNet-CNN-LSTM, and SeqNet-LSTM-CNN) per number of EAIs. The chart highlights the trend of the classification performance as the feature set-size grows in order to determine an optimal range for the number of EAIs.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

The Figure 4.3a shows a direct correlation between the AC and the number of EAIs from 8 to around 13, where it reaches its highest point. This means that using a moderate number of EAIs improve prediction by combining useful information from each one. After this peak, AC drops significantly, which suggests that adding more models does not always help and may even cause overlap or noise. The Accuracy lines follow stable trends with a low fluctuation between 10 and 12 number of EAIs, and shows best-performing range from 11 to 13 number of EAIs. In contrast, Accuracy is more variable when fewer EAIs are used, showing that small ensembles are more sensitive.

The Figure 4.3b presents the distribution of classification Accuracy as function of the number of EAIs from 8 to 16 among the Selected Models. Each box shows the spread of Accuracy scores across multiple ensemble runs. One can observed the direct correlation between the increment of the number of EAIs and the overall AC performance with narrower, and higher IQRs scores. This suggests that increasing EAIs diversity enhances predictive performance. Notably, the Accuracy distributions become tighter and higher between 10 and 15 number of EAIs, pointing to improved consistency and robustness in ensemble behaviour. Beyond 13, the gains in performance level off.

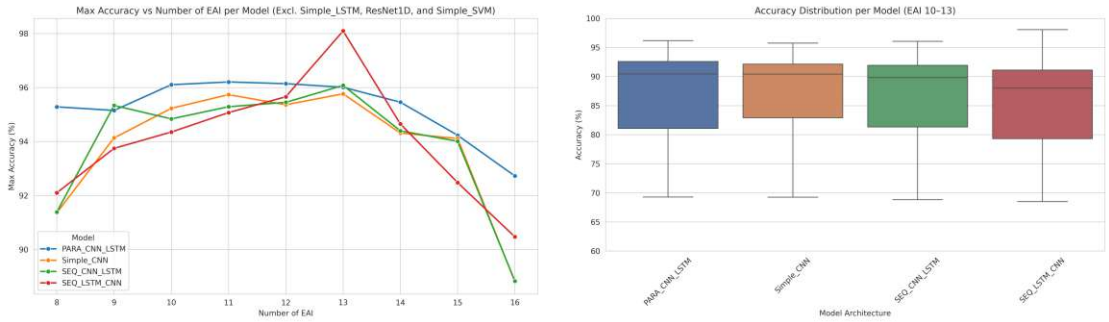
The most stable configurations with the highest performance are observed between 10 and 13 EAIs, where median Accuracy values approach or exceed 90%, and third Quartile (Q3) values consistently reach or surpass 92%. This range also exhibits narrower IQR, indicating low variability and high agreement among ensemble outputs. Maximum Accuracy values in this region often exceed 97.5%, demonstrating strong top-end performance. In contrast, smaller ensembles (8–9 EAIs) display wider IQRs and lower minimum values, suggesting greater sensitivity to input variability. Although using more than 13 EAIs maintains a relatively high average performance, it also introduces increased spread and more outliers, likely due to redundancy or conflicting contributions among the additional models.

Considering both subplots in Figure 4.3, the range of 10 to 13 EAIs offers the best trade-off between high Accuracy, low variance, and computational efficiency. Furthermore, the configuration with 13 EAIs achieves peak performance while maintaining robustness and computational efficiency, making it the most appropriated choice for the classification task of ecological soundscapes.

To complement the previous analysis, the line plot in Figure 4.4a illustrates the MAX Accuracy achieved by each of the four Selected Models architectures, such ParaNet-CNN-LSTM, Simple-CNN, SeqNet-CNN-LSTM, and SeqNet-LSTM-CNN, as a function of the number of EAIs, ranging from 8 to 16. Each curve represents the peak performance observed for a given architecture at each EAI ensemble size, offering a visual comparison of the performance among the Selected Models.

One can observe a similar trend among all the models. The AC improves from 8 to around 13 EAIs, where performance peaks, and then decreases for higher values. The highest AC values are consistently reached at 13 EAIs, indicating that this ensemble size provides the best trade-off between model performance, and EAIs diversity size. The observed decline beyond 13 EAIs suggests diminishing returns or possible conflicts

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN



(a) MAX AC by Number of EAIs per Model (b) AC Distribution per Model (10-13 EAIs)

Figure 4.4: Accuracy (AC) as a Function of the Number of EAIs per Model Excluding Simple-LSTM, Simple-SVM, and ResNet1D

introduced by adding more EAIs to the dataset.

Among all models, ParaNet-CNN-LSTM achieves the most stable and consistently high AC across the entire range. From 10 to 13 EAIs, its AC remains close above 96%. The line for this model is relatively flat and smooth in this region, showing low sensitivity to EAIs set-size changes and confirming strong robustness. This stability across multiple number of EAIs makes ParaNet-CNN-LSTM a reliable option under different configuration.

However, it is clear that SeqNet-LSTM-CNN reaches the single highest AC point at 13 EAIs, exceeding 98%, but its performance is less consistent. The curve shows larger fluctuations, particularly before 12 EAIs, and after the peak, with a sharp decline beyond 13 EAIs. This variability implies greater sensitivity to ensemble composition and limits its robustness for practical use where performance consistency is critical.

Nonetheless, Simple-CNN and SeqNet-CNN-LSTM demonstrate moderate performance. Both models reach AC values near 96% at their best, but they generally remain below ParaNet-CNN-LSTM and SeqNet-LSTM-CNN throughout the evaluated range. Their performance curves also show less pronounced peaks and more narrow margins of improvement, indicating a limited benefit from increasing the number of EAIs.

The boxplot in Figure 4.4b illustrates the distribution of Accuracy values obtained by the four Selected Models architectures across the ensemble range of 10 to 13 EAIs. In the Figures 4.4b and 4.5 one can observed that the ParaNet-CNN-LSTM model achieves consistently the highest median and third Quartile (Q3) AC values, with a narrow interquartile range concentrated in the high-performance region. Its relatively short lower whiskers indicate consistent predictive behaviour and limited susceptibility to performance drops. In comparison, Simple-CNN displays an even more compact Interquartile Range (IQR), but with lower maximum and median AC values. This reflects greater robustness but a lower overall predictive ceiling. The SeqNet-CNN-LSTM and SeqNet-LSTM-CNN models exhibit wider distributions and slightly lower stability than Simple-CNN, and their performance remains below that of the top-performing ParaNet-CNN-LSTM hybrid.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

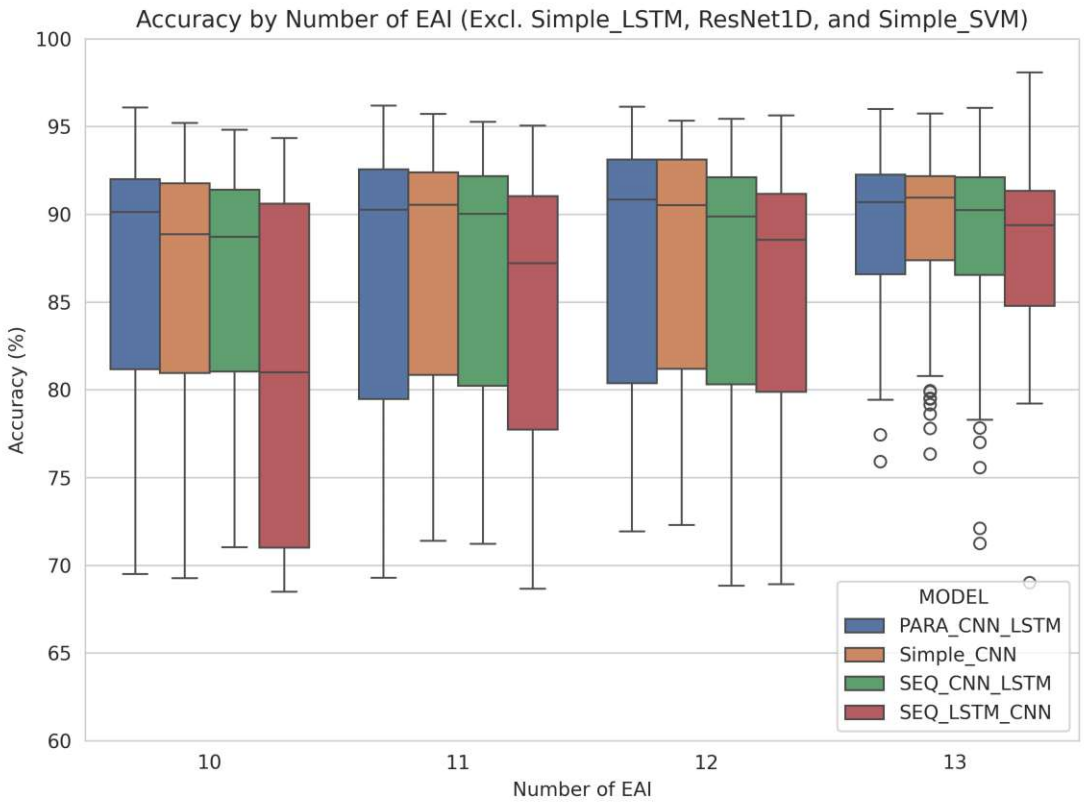


Figure 4.5: AC Distribution by Number of EAIs per Model Excluding Simple-LSTM, Simple-SVM, and ResNet1D

Overall, as it is presented in Figure 4.5, the ParaNet-CNN-LSTM stands out as the most effective and reliable model architecture for this specific ecological soundscape classification task. It consistently delivers one the highest Maximum (MAX), third Quartile (Q3), and median Accuracy across the critical ensemble size-range of 10 to 13 EAIs, while maintaining low variability and a compact Accuracy distribution. This robustness is evidenced by its stable performance, short lower whiskers, and minimal extreme outliers that proposes resilience to fluctuations across evaluation splits. In comparison, the SeqNet-LSTM-CNN model occasionally reach competitive peak values but suffer from greater sensitivity and broader dispersion, reducing their overall reliability. Likewise, Simple-CNN and SeqNet-CNN-LSTM show moderate performance and lack the stability and peak Accuracy performance. Taken together, these characteristics validate ParaNet-CNN-LSTM as the most suitable choice, offering the best trade-off between predictive strength, stability, and generalisation for the classification task.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

4.2.4 Individual Contribution of Each EAI to Model Performance

For this experiment, we evaluated all 511 unique feature combinations across the four Selected Models (ParaNet-CNN-LSTM, Simple-CNN, SeqNet-CNN-LSTM, and SeqNet-LSTM-CNN), resulting in a total of 2044 model runs. Each feature set was constructed by appending the fixed set of 7 baseline features to an unique combination of Ecological Acoustic Indices (EAIs), using the same generation process described in Section 4.1. Then, we one-hot encoded the EAIs in the results-dataset, with each activated EAI assigned the corresponding Accuracy (AC) from its respective prediction run, as illustrated in each subplot of Figure 4.6.

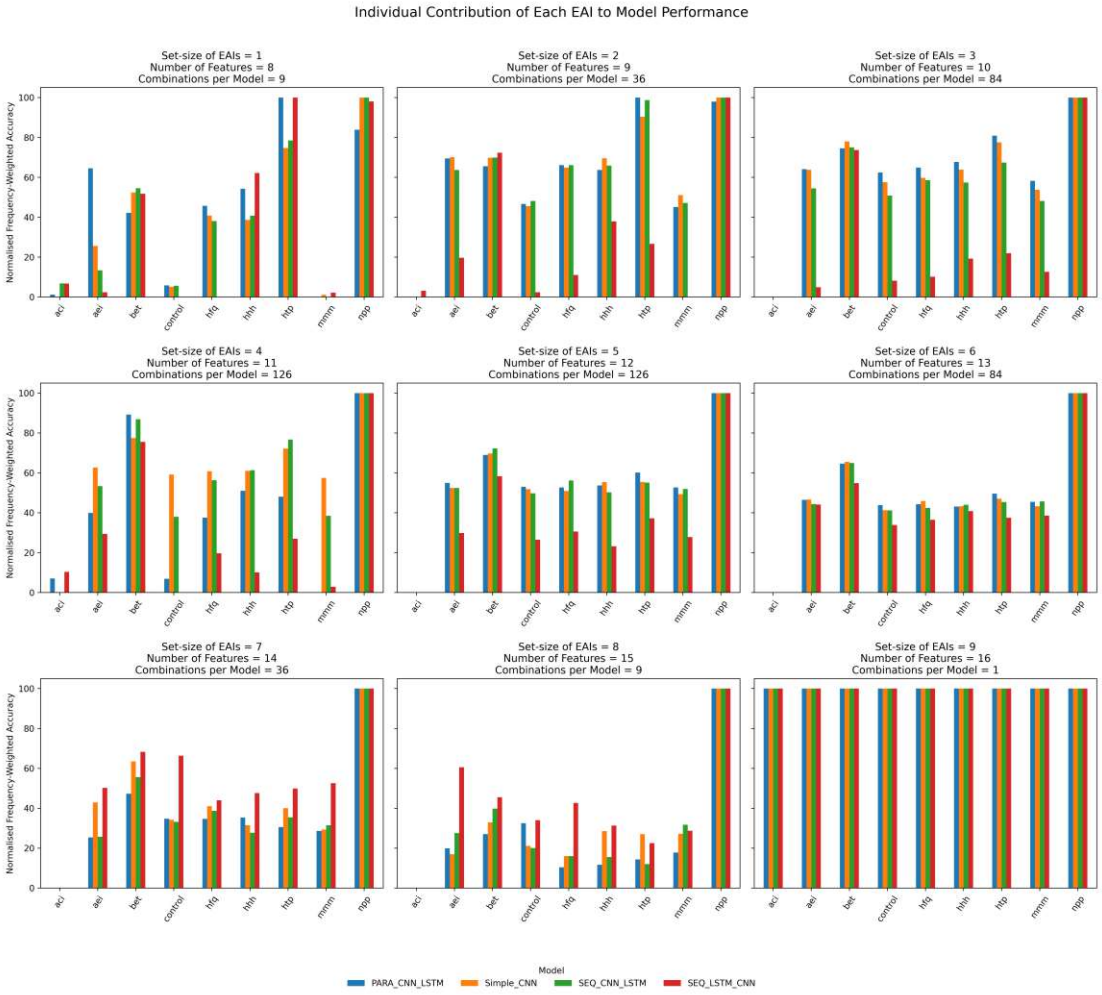


Figure 4.6: Individual EAI Contribution to Model Performance

It is important to note that we assumed statistical independence among the EAIs. Additionally, we used the Maximum (MAX) amplitude of the signal from the curated

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

dataset described in Section 2.3.2 as a reference threshold. This feature was relabelled as `control` and used as a baseline for comparison, allowing us to assess whether individual EAIs contributed more or less than this threshold to the overall classification performance. Lower NFWA values indicate weak individual contributions, while higher NFWA values reflect a stronger positive influence on the AC performance.

The Figure 4.6 illustrates the individual contribution of each EAI to model performance, across all the set-sizes. The number of distinct EAIs used per feature set, ranging from 1 to 9. The vertical axis represents the Min-Max Normalised Aggregated Frequency-Weighted Accuracy (NFWA), scaled from 0 to 100. Each subplot corresponds to a specific set-size and shows the NFWA contributions for each EAI across the four Deep Learning Selected Models (ParaNet-CNN-LSTM, Simple-CNN, SeqNet-CNN-LSTM, and SeqNet-LSTM-CNN). The number of feature combinations varies depending on the set-size, as detailed in Table 4.1.

In the early stages, from set-size 1 to 3, there is a clear upward trend in NFWA. Combining two or three EAIs already brings notable improvements compared to single-feature sets. In particular, EAIs such as NPP, HTP, and BET begin to stand out with higher contributions, whereas others like ACI, AEI, MMM, and `control` show lower individual impact. These smaller ensembles highlight the early synergistic effects of specific EAIs when paired.

Between set-size 4 and 5, the trend of increasing NFWA continues more strongly. EAIs like NPP, HTP, and BET consistently show significant influence across models. In addition, features such HFQ, and AEI also show positive impact-gains in the performance. These configurations demonstrate a more balanced ensemble with higher entropy, offering a strong trade-off between feature richness and model robustness.

At set-size 6, the most stable and balanced ensemble is observed across all the Selected Models. The contribution scores of individual EAIs are nearly equal, indicating a well-distributed and synergistic feature combination. It also means that this is the Maximum optimal set-size. Despite this balance, subtle variations still reveal that EAIs such as NPP, BET, HTP, AEI, HHH, end even MMM remain slightly more impactful. This ensemble size achieves high Accuracy without dominance by any single-feature, apart from the clear NPP domination across all the set-sizes, suggesting it is a robust configuration with reduced risk of overfitting.

For set-sizes 7 and 8, the Accuracy exhibits a slight decline across most EAIs and Selected Models, with the exception of the SeqNet-LSTM-CNN, which shows modest gains. Nonetheless, the overall performance of all Selected Models decreases within this set-size range, as detailed in Section 4.2.3 and illustrated in Figure 4.3a.

At set-size 9, all EAIs are included in a single combination, leading to an unique-single MAX AC value. However, this result is limited in interpretability, as there is only one configuration and it likely overfits the training data.

To enable a more detailed analysis, we focused on a closer inspection of the critical

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

feature set-size range from 10 to 13 as identified in Section 4.2.3. This range is further illustrated in Figures 4.7 and 4.8, allowing a focused comparison across the Ecological Acoustic Indices (EAIs) and the Selected Models.

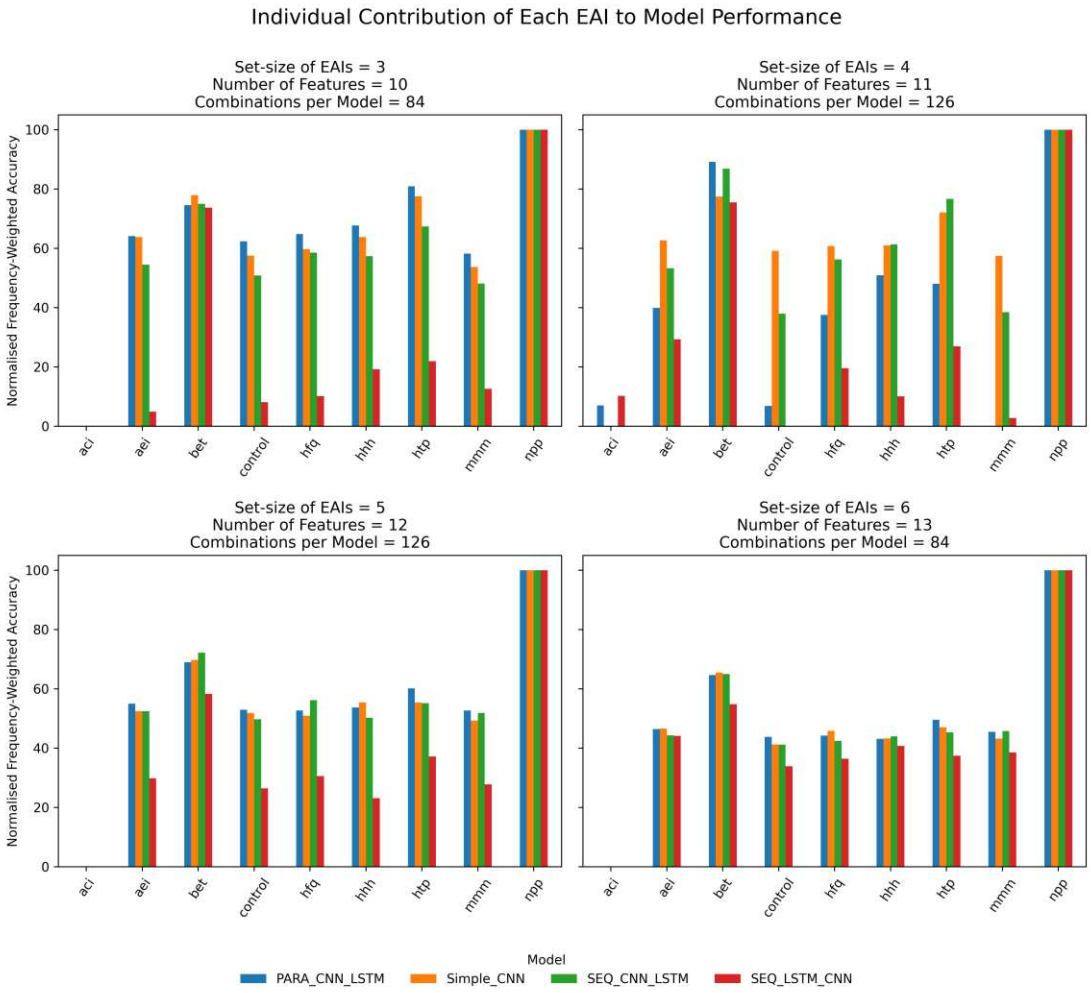


Figure 4.7: Individual EAI Contribution for 10-13 Number of Features

In Figure 4.7, it can be observed that within the set-size range from 10 to 13, EAIs such as NPP, BET, HTP, AEI, HFQ, and HHH consistently demonstrate strong contributions across both set-sizes and models, identifying them as the most impactful features. The NFWA increases steadily from set-size 3 to 5, beyond which the performance begins to plateau. This trend suggests that ensembles comprising approximately five 5 well-selected EAIs offer an effective trade-off between feature diversity and predictive performance.

Furthermore, Figure 4.7 shows that the ParaNet-CNN-LSTM consistently outperforms the other Selected Models, achieving the highest NFWA, particularly at higher set-

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

sizes. This highlights its capacity to better leverage additional features, although other CNN-LSTM variants also benefit from larger ensembles. These results underscore the critical role of model architecture in determining how effectively feature set-size and feature-selection interact to enhance predictive performance.

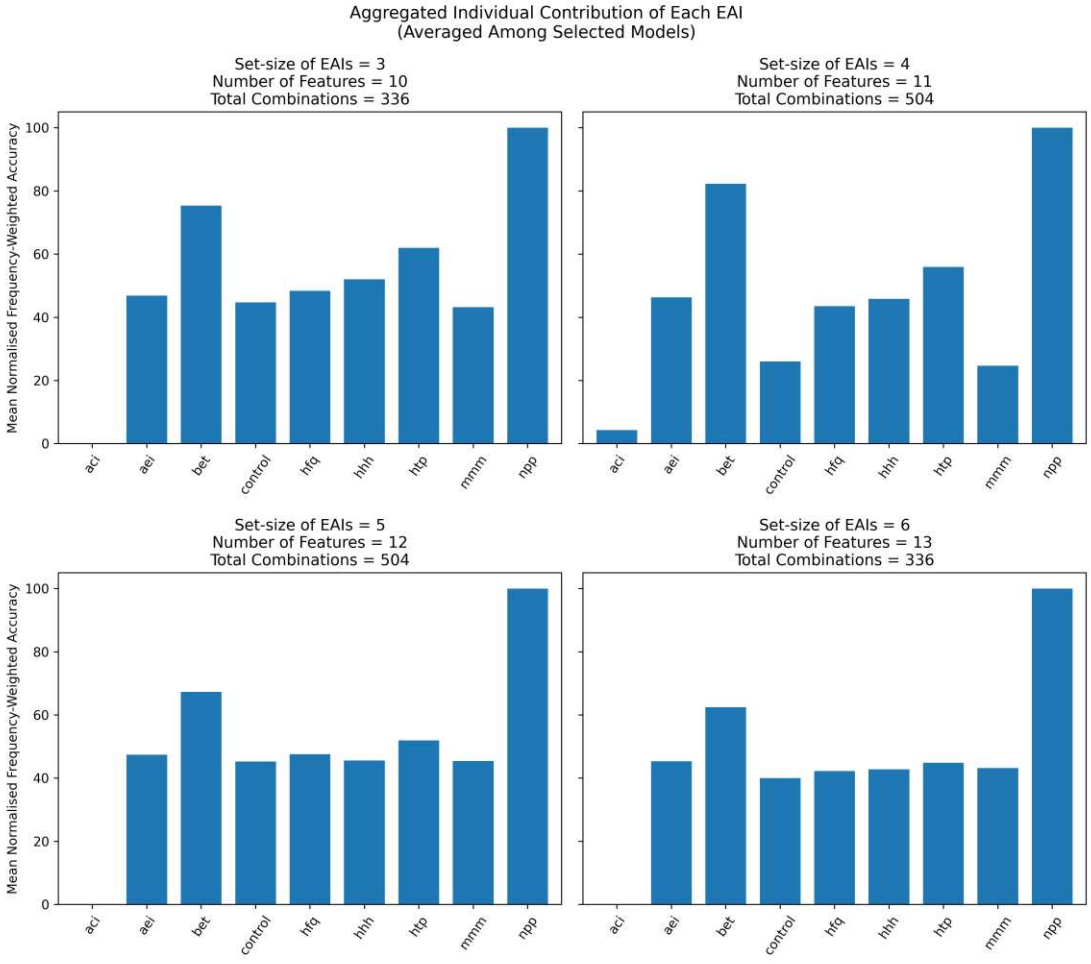


Figure 4.8: AGG Individual EAI Contribution Among Selected Models for 10-13 Number of Features

To summarise the insights, the Figure 4.8 presents the Aggregated individual contributions of each EAI, averaged across the Selected Models, for set-sizes ranging from 10 to 13 total features (corresponding to 3 to 6 EAIs). Each subplot displays the mean Normalised Aggregated Frequency-Weighted Accuracy per EAI, capturing how much each individual feature contributes to overall classification performance within its respective set-size.

Across all set-sizes, the NPP feature consistently achieves the highest NFWA, establishing

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

it as the most influential EAI in the ensemble. Features such as BET, HTP, AEI, and HFQ also demonstrate consistently strong contributions, particularly for set-sizes of 4 and 5. In contrast, features like MMM and HHH generally exhibit weak performance, often yielding NFWA scores comparable to or lower than the `control` index, which retains a low value, thereby supporting its role as a neutral reference threshold.

As the number of features increases, the overall contribution distribution becomes more even across the EAIs, indicating that larger ensembles may promote redundancy or complementary interactions. Notably, the saturation of NFWA at higher set-sizes supports the hypothesis that a well-chosen subset of 5 EAIs may achieve near-optimal performance with fewer features.

In conclusion, using a feature set-size of 12 total features (corresponding to 5 EAIs) gives the most robust, and accurate results, while adding more features brings little or no improvement. The Ecological Acoustic Indices (EAIs) that contribute most with model performance are Number of Peaks (NPP), Bioacoustic Index (BET), Temporal Entropy (HTP), Acoustic Evenness Index (AEI), and Frequency Entropy (HFQ).

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 5

Conclusion

In conclusion, this thesis synthesises the key insights and summarises the obtained results. The main contributions and findings are then outlined, followed by a discussion that reviews and answers the Research Questions. We also reflect on the limitations encountered in this study, clarifying certain approaches that can not be generalised. Finally, we propose directions for future research, encapsulating the proposal of the **F**ramew**O**rk for featu**R**e **E**xtraction, vi**S**ualisation, and classifica**T**ion of Soundscapes.

5.1 Summary of Contributions

In this section, we summarise and discuss the three (3) Research Questions (RQs) introduced in Chapter 1. Each RQ has been addressed through tangible contributions that guided the execution of this thesis, with answers driven by the design, implementation, and evaluation of the proposed FOREST.

RQ1. What are the appropriate data pre-processing and Data Augmentation techniques that can be applied to raw audio recordings (*soundscapes*) to perform model training, and impact predictive performance?

We establish a structured data-processing workflow, presented in Section 2.3, that transforms raw soundscape recordings into a curated dataset suitable for model training. The pipeline converts 57 audio records of 30 minutes each into overlapping 6-second frames and applies several data transformations and Data Augmentation techniques, as detailed in Section 2.1, including wave enveloping, noise reduction, spectral transformations, framing with Hanning windows, and Fast Fourier Transform to separate frequency bands into distinct vectors. For each band and frame, we extract five (5) statistical scalars and eleven (11) established Ecological Acoustic Indices, such as ACI, ACA, ADI, BET,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

MMM, NPP, HFQ, HTP, HHH, AEI, and NDSI. This process yields feature tensors that integrate spectral, temporal, and ecological information.

The curated dataset presented in Section 2.3.2, with a final shape of 249,660 × 6,016, is derived from raw soundscapes across four (4) labelled Regions and provides a consolidated representation for feasible training and evaluation of the proposed DL-based models.

RQ2. What are the most relevant characteristics of the soundscapes among the ecosystem Regions (Reference Forest, Pasture, Natural Regeneration, and Plantation)?

Building upon of the curated dataset presented in Section 2.3.2, an evaluation framework is introduced in Section 4.1 to analyse the influence of individual features on classification performance. From the total pool of 16 features, a fixed subset of seven (7) is combined with all non-empty subsets of nine (9) variable EAIs, yielding $2^9 - 1 = 511$ unique combinations. Each DL-based model architecture described in Section 3.2 is trained and evaluated 511 times, resulting in $7 \times 511 = 3577$ experiment runs. This systematic design enables a robust assessment of both the number and the specific contribution of each EAI to the predictive performance per model.

The results presented in Section 4.2.4 highlight a consistent ranking of feature importance proven reliability on the selected EAIs. Among the eleven (11) Ecological Acoustic Indices, Number of Peaks (NPP) emerges as the most influential, while Bioacoustic Index (BET), Temporal Entropy (HTP), Acoustic Evenness Index (AEI), and Frequency Entropy (HFQ) also provide strong and reliable contributions. In contrast, Temporal Median (MMM) and Combined Entropy (HHH) show limited predictive value, often performing near the neutral `control` index. Furthermore, the findings demonstrate that a feature set-size of twelve (12) total features, seven (7) fixed statistical features plus five (5) Ecological Acoustic Indices, achieves the most accurate and robust results, while adding additional indices yields overfitting, and poorer or none model performance.

The spidernet charts presented in Section 2.3.3 show the characteristics of the five (5) selected EAIs across the Regions. Distinct patterns can be observed among them. In particular, the Plantation ecosystem exhibits low values for most EAIs, except for a moderate value of HFQ. In contrast, the Pasture ecosystem shows high values for BET and HTP, alongside minimal values for NPP and HFQ. Natural Regeneration records the lowest AEI, with low BET, moderate to high NPP and HTP, and the highest HFQ. Finally, the Reference Forest displays consistently high values across most indices, with peaks in NPP and AEI, high HFQ and BET, but relatively low HTP.

RQ3. What is an appropriate Deep Learning model architecture to learn the statistical patterns of soundscapes and accurately classify audio signals among the different environmental ecosystem Regions?

In Section 3.2, we present three (3) DL-based models, namely ParaNet-CNN-LSTM, SeqNet-CNN-LSTM, and SeqNet-LSTM-CNN, which combine CNN and LSTM layers. This family of hybrid architectures is capable to capture both spectral patterns and

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

temporal dependencies from soundscapes. Next, Section 3.2.4 introduces the baseline models for benchmarking, including Simple-SVM, Simple-CNN, Simple-LSTM, and a customised ResNet1D.

The evaluation framework presented in Section 4.1 not only analyses the contribution of individual EAIs but also enables a comparative assessment of different model architectures. Each of the seven (7) DL-based models is trained and evaluated 511 times, yielding a total of 3,577 experimental runs. This systematic setup provides a robust method for identifying the most appropriate DL-based architecture for this specific environmental soundscape classification task.

The evaluation results presented in Section 4.2.3 indicate that the ParaNet-CNN-LSTM consistently achieves the most reliable performance across varying numbers of EAIs, particularly in the range of 10 to 13, where median Accuracy values exceed 90% and maximum values surpass 97.5%. Its stable Accuracy (AC) distribution, consistently achieving the highest median and third Quartile AC values, narrow interquartile range, relatively short lower whiskers, and resilience to outliers demonstrate strong robustness and generalisation. While the SeqNet-LSTM-CNN occasionally reaches competitive peak performance above 98%, it shows higher variability and sensitivity to feature set-size.

These findings validate the ParaNet-CNN-LSTM as the most appropriate architecture for this specific ecological soundscape classification task. It offers the best trade-off between predictive capabilities, stability, and generalisation, establishing a reliable methodological foundation for the FOREST.

The **F**ramew**O**rk for featu**R**e **E**xtraction, vi**S**ualisation, and classifica**T**ion of Soundscapes (FOREST) consolidates these contributions into a modular Python-based system. It unifies preprocessing, dataset curation, feature extraction, visualisation, predictive modelling, and evaluation within a single framework. The system is made available as an open-source, web-based application at `www.soundforest.app`.

5.2 Discussion of Limitations

The main limitations of this study concern dataset characteristics, temporality assumptions, and scope of applicability. Recognising these constraints is essential for interpreting the contributions, and guiding future extensions of the FOREST framework.

Imbalanced Dataset. The dataset used in this study is highly imbalanced, as not all Regions contain the same number of recorded soundscapes. To address this, a stratification strategy is applied to ensure that class proportions remain consistent across the training, testing, and evaluation splits as presented in Section 3.3.1. This enables a more reliable assessment of model performance within the scope of the available dataset. However, such stratification is not optimal when extending the dataset or aiming to generalise the approach to a larger number of classes.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Atemporality. The temporal coverage of the dataset assumes recordings were made in parallel at the same time. It is important to note that usual animal activity varies substantially across daytime hours and seasonal cycles, reflecting natural behavioural patterns. For this reason, the dataset relies on the expertise of biologists and environmental conservation specialists, who previously annotated the soundscapes dataset.

Absence of Metadata. Information such as the geo-location of soundscapes, and complementary ecological ecosystem-characteristics (e.g., vegetation type, annual precipitation, or canopy height) would have been highly beneficial. Such metadata could enhance the potential of multi-modal predictive models by incorporating environmental context alongside the acoustic features.

Limited to the PSA Program. The scope of the soundscapes is restricted to the PSA program in Costa Rica, meaning that the FOREST framework cannot yet be generalised beyond this domain-specific dataset.

Regardless of these limitations, the scientific methodology implemented in this thesis follows a rigorous foundation. We aim that the proposed FOREST could be applied to larger and more diverse PAM-based projects in distinct geographical regions.

5.3 Further Research

There are several directions in which future work could extend this thesis. Some experiments yielded promising results, showing that the proposed hybrid DL-based model architectures similar to FOREST can be further refined and enhanced. At the same time, others revealed limitations, and pointing to alternative strategies. Beyond these, many research gaps remain outside the scope of this work, offering unexplored opportunities for advancing environmental soundscape classification with Deep Learning.

Data Augmentation with MIR Techniques. The promising results obtained with Ecological Acoustic Indices and basic Music Information Retrieval (MIR)-based Data Augmentations indicate strong potential for exploring cross-domain approaches towards audio classification. Extend Data Augmentation strategies inspired by Music Information Retrieval to investigate new paradigms of sound transformation and processing [133]. For instance, similarities may be drawn between musical genres and ecological soundscapes, or between the calls of specific animal species and the sounds of musical instruments. Such analogies could provide novel perspectives on the relationship between human-generated sound and natural ecosystems, fostering more harmonious interactions with hyper-diverse environments such as rain-tropical forests.

Dataset Expansion Across Ecosystems and Time. Extend the dataset beyond the Payment for Environmental Services program in Costa Rica to include other ecosystems

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

and longer temporal spans. The current study assumes recordings from same daytime hours and seasonal windows to maintain homogeneity and simplify comparisons. Future research should deepen this scope by incorporating diverse geographic regions, additional ecosystem categories, and richer temporal variability [6], [12], [20], [26], [29], [30]. Moreover, more detailed metadata could be included, such as annual precipitation, enhanced vegetation indices, human modification levels, canopy height, and elevation [6], [8], [13], [14], [17], [20], [23], [24]. Such ecological covariates would enhance the capacity to stablish a bio-diversity index based on ecological soundscapes using Passive Acoustic Monitoring.

Robust Acoustic Scene Classification Framework. Develop more robust classification models inspired by the annual Detection and Classification of Acoustic Scenes and Events Acoustic Scene Classification challenges [118]–[122], [130]. These tasks highlight critical issues such as device mismatch, low-complexity constraints, and data efficiency, which are equally relevant to ecological monitoring. Adapting methods from ASC research, such as device-specific modelling, cross-device generalisation, or data-efficient pre-training could significantly improve the robustness and scalability of the FOREST. In addition, experimenting with Transformer-based architectures [100], such as UVCGAN [181], may extend the capabilities of eco-acoustic classification models.

Knowledge Graph for Unsupervised Subsampling. When extending the analysis to long-duration recordings spanning several hours, days, or even weeks, as is common practice in ecological Passive Acoustic Monitoring [2], [6], [26], [53], an important research direction is the integration of Knowledge Graph (KG)-based strategies similar to Pillar II (Decentralised and Strategic Decision Making) and Pillar III (Enhanced Human–AI Interaction) of the INTEND project [182]. These approaches could support unsupervised subsampling by automatically identifying the most informative temporal segments [6], [26]–[28], such as periods marked by peaks in specific Ecological Acoustic Indices. Beyond reducing computational bottlenecks while applying Cognitive Computing Continuum [183], these methods would also enhance human–AI collaboration, providing ecologists with more intuitive tools to interact with automated systems, and reducing the time-intensive limitations of manually screening extended recordings.

Open-Source Framework Development & Public Release. To encapsulate these contributions and support adoption in ecological conservation, the FOREST could be developed into a more robust open-source Python framework, comparable to REAL [70], CMIP [184], [185], Scikit-MAAD [43], and OpenSoundscape [98]. Future efforts could focus on refining, documenting, usability, and releasing it as a Python package accessible to practitioners with minimal coding expertise.

The current **FOREST** implementation remains at an early stage. Further research could significantly support large-scale Passive Acoustic Monitoring efforts in biodiversity conservation, establish a foundation for doctoral-level research, and foster long-term community-driven collaboration among prestigious research institutions worldwide.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

APPENDIX A

DSRM Artifacts

The aim of this DSRM artifact list is to summarise the specific contributions delivered at each stage of the research project that are directly relevant to this work.

A.1 Delivered Artifacts

Artifact: Master Thesis Proposal

Status: Completed

Format: Document

Phase: Problem Identification & Motivation, Define Objectives

Activities: Initial meetings with stakeholders, shadowing existing data workflows & methods

Methods: Software Requirements Engineering (main), SLR (main)

Acceptance Criteria: Approval from external stakeholders and supervisors, master thesis proposal acceptance

Stakeholders: Supervisors

Tentative Dates: 24.06.2024 – 24.12.2024

Estimated Effort: 5 months

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Artifact: Preliminary Table of Content

Status: Completed

Format: Document

Phase: Problem Identification & Motivation, Define Objectives

Activities: Initial meetings with stakeholders, shadowing existing data workflows & methods, and SLR

Methods: Software Requirements Engineering (main), SLR (main)

Acceptance Criteria: Approval from external stakeholders and supervisors, master thesis proposal acceptance

Stakeholders: Supervisors

Tentative Dates: 01.01.2025 – 02.02.2025

Estimated Effort: 1 month

Artifact: Preliminary Project Plan Overview

Status: Completed

Format: Document

Phase: Define Objectives, Design & Development, Artifact Development

Activities: Project plan overview design

Methods: SCRUM (main), CRISP-DM (main), DSRM (secondary)

Acceptance Criteria: Evidence of addressing research questions, alignment with research deadlines, supervisor approval, research methods seminar paper acceptance

Stakeholders: Supervisors, Research Methods Seminar Professor

Tentative Dates: 01.01.2025 – 02.02.2025

Estimated Effort: 1 month

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Artifact: SLR Protocol (Appendix B)

Status: Completed

Format: Document

Phase: Problem Identification & Motivation, Define Objectives

Activities: Initial meetings with stakeholders, shadowing existing data workflows & methods, SLR.

Methods: SLR (main)

Acceptance Criteria: Research methods seminar paper acceptance

Stakeholders: Supervisors, Research Methods Seminar Professor

Tentative Dates: 01.01.2025 – 02.02.2025

Estimated Effort: 1 month

Artifact: Prototype Development

Status: Completed

Format: Working Software

Location: Git Repository (`https://github.com/carlosvargas9103/soundforest`)

Phase: Define Objectives, Design & Development

Activities: Software development

Methods: SCRUM (main), SLR (main), CRISP-DM (secondary)

Acceptance Criteria: Supervisors approval, preprocessing implemented, bioacoustic methods implemented, initial model training and testing

Stakeholders: Supervisor

Tentative Dates: 01.11.2024 – 20.05.2025

Estimated Effort: 6 months

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Artifact: FOREST - Application Development

Status: Completed

Format: Working Software

Location: Public Website (`https://soundforest.app/`)

Phase: Demonstration & Implementation

Activities: SCRUM with bi-weekly sprints, modular implementation

Methods: SCRUM (main), CRISP-DM (main), Clean Architecture (secondary), SLR (secondary)

Acceptance Criteria: Supervisors approval, evidence of addressing Research Questions

Stakeholders: Supervisors

Tentative Dates: 11.11.2024 – 09.09.2025

Estimated Effort: 10 months

Artifact: FOREST - Preprocessing Tasks for Time Series Data

Status: Completed

Format: Dataset

Location: Zenodo Repository (`https://doi.org/10.5281/zenodo.17070625`)

Phase: Demonstration & Implementation

Activities: SCRUM with bi-weekly sprints

Methods: SCRUM (main), CRISP-DM (main), Clean Architecture (secondary), SLR (secondary)

Acceptance Criteria: Supervisors approval, evidence of addressing research questions

Stakeholders: Supervisors

Tentative Dates: 01.11.2024 – 01.12.2024

Estimated Effort: 2 months

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Artifact: FOREST - Apply Bioacoustic Metrics

Status: Completed

Format: Dataset

Location: Zenodo Repository (`https://doi.org/10.5281/zenodo.17070625`)

Phase: Demonstration & Implementation

Activities: SCRUM with bi-weekly sprints

Methods: SCRUM (main), CRISP-DM (main), SLR (main)

Acceptance Criteria: Supervisors approval, evidence of addressing research questions

Stakeholders: Supervisors

Tentative Dates: 11.11.2024 – 10.01.2025

Estimated Effort: 2 months

Artifact: FOREST - Curated Dataset

Status: Completed

Format: Dataset

Location: Zenodo Repository (`https://doi.org/10.5281/zenodo.17070625`)

Phase: Demonstration & Implementation

Activities: SCRUM with bi-weekly sprints

Methods: CRISP-DM (main), SLR (main)

Acceptance Criteria: Supervisors approval, able to be used to train the models, evidence of addressing research questions

Stakeholders: Supervisors

Tentative Dates: 11.11.2024 – 03.03.2025

Estimated Effort: 3 months

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Artifact: Model Architecture

Status: Completed

Format: PyTorch Objects

Location: Zenodo Repository (`https://doi.org/10.5281/zenodo.17024884`)

Phase: Demonstration & Implementation, Evaluation

Activities: SCRUM with bi-weekly sprints

Methods: SCRUM (main), CRISP-DM (main), SLR (main)

Acceptance Criteria: Supervisors approval, evidence of decent performance time, and prediction scores.

Stakeholders: Supervisors

Tentative Dates: 12.12.2024 – 13.03.2025

Estimated Effort: 3 months

Artifact: Model Training & Testing and Evaluation

Status: Completed

Format: Performance Results JSON

Location: Zenodo Repository (`https://doi.org/10.5281/zenodo.17024884`)

Phase: Demonstration & Implementation, Evaluation

Activities: SCRUM with bi-weekly sprints

Methods: CRISP-DM (main)

Acceptance Criteria: Supervisors approval, evidence of decent performance time, and prediction scores.

Stakeholders: Supervisors

Tentative Dates: 12.12.2024 – 13.03.2025

Estimated Effort: 3 months

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Artifact: Models Performance Comparison with SOTA Models

Status: Completed

Format: Performance Results JSON

Location: Zenodo Repository (`https://doi.org/10.5281/zenodo.17024884`)

Phase: Evaluation

Activities: SCRUM with bi-weekly sprints

Methods: SCRUM (main), CRISP-DM (main), SLR (main)

Acceptance Criteria: Supervisors approval, evidence of decent performance time, and prediction scores.

Stakeholders: Supervisors

Tentative Dates: 12.02.2025 – 05.05.2025

Estimated Effort: 3 months

Artifact: Master Thesis Manuscript and Presentations

Status: Completed

Format: Document

Phase: Communication & Research Contribution

Activities: Scientific writing, SLR

Methods: Scientific Writing (main), SLR (main), CRISP-DM (secondary)

Acceptance Criteria: Supervisors approval, evidence of addressing research questions

Stakeholders: TU Wien Informatics Decanature

Tentative Dates: 11.11.2024 – 09.09.2025

Estimated Effort: 10 months

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

APPENDIX B

SLR Artifacts

The aim of this SLR is to systematically collect, evaluate, and summarise existing contributions across specific research topics that are directly relevant to this work.

B.1 SLR Protocol

Before conducting a Systematic Literature Review (SLR), it is essential to establish the motivation and clarify its necessity within the research project. The value of an SLR lies in assessing whether existing reviews already cover the topic of interest, and in identifying how a new review can provide additional insights.

The search strategy for the SLR protocol involves selecting keywords, databases, type of documents, time-frame, language, criteria of inclusion-exclusion, journal quality, number of citations, and number of authors. These decisions directly affect the quality of the review. However, the search terms can vary to refine the SLR process, and identify more relevant scientific insights from the literature on the specified topics. Also, the selection process should be transparent, justified, and replicable to ensure credibility. While conducting an adequate review we can ensure clarity and rigour throughout the process. Therefore, the search strategy presented in Section B.1.1 has been followed in this master thesis.

B.1.1 Search strategy Design

The search strategy process will be based on the search of **"Topics"** among certain domains or **"Groups"**. The combination of these Topics, and Groups will serve to the Articles Extraction Process. After some of the most relevant Articles for the Research Questions (RQs) are selected, the further steps will assure that only the most influential Articles will be used for the Master Thesis Research.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

STEP 0 - Definition of the Topics and Groups

Topics: Based on the RQs.

RQ1: Preprocessing, Data Transformation, Data Augmentation (DA), Music Information Retrieval (MIR), Ecological Acoustic Indices (EAIs)

RQ2: Data Exploratory Analysis (DEA), Cluster Analysis, Feature Extraction (FE), Factor Analysis (FA), Data Visualisation

RQ3: Machine Learning (ML) Framework, Modelling, Model Architecture, Classification Model, Predictive Model, Deep Learning (DL) Model, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Model Evaluation, Performance Evaluation, Classification, Datasets

Groups: Domain Specific.

RQ1-3: Time Series, Audio Signals, Audio Records, Acoustics, Bioacoustics, Soundscapes, Environmental Audio

STEP 1 - Extraction of Papers from Library Databases

Data Bases: The specific bases that are relevant to the Research questions. The most relevant Databases that will be used for this research project are: Österreichische Dissertationsdatenbank, IEEE/IET Electronic Library (IEL), ACM Digital Library, arXiv, ScienceDirect, Springer Link, Scopus.

Document Type: The type of document to query. For the most of the search queries they will be either Scientific **Articles and/or Books.**

Timespan: The search will be filtered by only the items that are not older than 10 years (2015 - 2025). This is due to the dynamist of the research field. The goal is to retrieve the most relevant advances on the field from current research outcomes.

Language: English.

Query: By using the **Topics** in truncated combinations with the **Groups** as search strings.

STEP 2 - Identification of Relevant Article

Method: Scientific reading of the Abstract and the full article.

Acceptance Criteria: There will be three (3) main requirements.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

- The Article addresses, at least partially, one of the RQ.
- The Article utilises Time Series Data, or Audio Records.
- Involvement of at least four (4) Authors.

STEP 3 - Manually Tracking Citations For Additional Relevant Articles

Method: Internet Information Search for citation tracking.

Acceptance Criteria: There will be two (2) main requirements.

- Same Criteria than for STEP 2.
- Highly ranked journals.

STEP 4 - Identification of Most Influential Articles

Method: Scientific reading of the full article.

Acceptance Criteria: There will be Three (3) main requirements.

- The Article has been recently published.
- The Article addresses more than one of the RQs.
- The number of citations per year.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

List of Figures

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

List of Tables

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

List of Algorithms

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Acronyms

ParaNet-CNN-LSTM, Simple-CNN, SeqNet-CNN-LSTM, and SeqNet-LSTM-CNN Selected Models. 65–67, 69–72, 93

AC Accuracy. xi, xiii, 31, 32, 37, 46, 53–57, 59, 61–70, 72, 77, 93, 101

ACA Alternative Acoustic Complexity Index. 19, 26, 75

ACI Acoustic Complexity Index. 17–19, 26, 60, 70, 75

Adam Adaptive Moment Estimation. 51, 53, 97

ADI Acoustic Diversity Index. 20, 21, 25, 26, 75

AEI Acoustic Evenness Index. xi, xiii, 25, 26, 30, 60, 70, 71, 73, 76

AGG Aggregated. 29, 54, 55, 61–63, 65, 70, 72, 93, 101

AI Artificial Intelligence. 79

ASC Acoustic Scene Classification. 79

ASC Austrian Scientific Computing. vii, ix, 31, 50

BET Bioacoustic Index. xi, xiii, 21, 26, 29, 30, 60, 70, 71, 73, 75, 76

Bi-LSTM Bi-directional Long Short-Term Memory. 37, 100

CCC Cognitive Computing Continuum. 79

CMIP Coupled Model Intercomparison Project. 79

CNN Convolutional Neural Network. xi, xiii, 3, 5, 8, 31, 32, 34, 36–38, 40, 42–49, 52, 53, 55–57, 61, 63–70, 72, 76, 77, 90, 99–102

COLFUTURO Fundación para el Futuro de Colombia. vii, ix

CRISP-DM Cross-Industry Standard Process for Data Mining. 3, 4, 6, 7, 82–87

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

CRNN Convolutional Recurrent Neural Networks. 35

D Dimensional. xi, xiii, 38–40, 43–45, 53, 56, 77

DA Data Augmentation. 3–5, 7, 9, 12, 75, 78, 90

DCASE Detection and Classification of Acoustic Scenes and Events. 79

DEA Data Exploratory Analysis. 5, 6, 90

DL Deep Learning. xi–xiii, xv, 1, 3, 4, 7–9, 31–38, 40, 42–46, 48, 50, 52, 54, 56–58, 61, 70, 76–78, 90

DSRM Design Science Research Method. xvi, 3–7, 81, 82, 84, 86

EAI Ecological Acoustic Indice. xi–xiii, xv, 2, 5–9, 11, 12, 15, 17, 19, 21, 23, 25, 26, 29, 30, 59–73, 75–79, 90, 93, 95

EECP-CBL Electric Energy Consumption Prediction model utilising the combination of Convolutional Neural Network and Bi-directional Long Short-Term Memory. 37

EPFL École Polytechnique Fédérale de Lausanne. i, iii, vii, ix, 31, 50

ETH Zürich Eidgenössische Technische Hochschule Zürich. i, iii

FA Factor Analysis. 5, 90

FC Fully Connected. 34, 41–43, 45, 47, 48

FE Feature Extraction. 3, 5, 7, 9, 17, 90

FFT Fast Fourier Transform. 14, 16, 23, 26, 75

FN False Negatives. 54

FNN Feedforward Neural Network. 33

FOREST **F**ramew**O**rk for featu**R**e **E**xtraction, vi**S**ualisation, and classifica**T**ion of Soundscapes. xi, xiii, 3, 6–8, 31, 36, 48, 52, 53, 75, 77–79

FP False Positives. 54

GPU Graphics Processing Unit. 31, 50, 52, 54

GRU Gated Recurrent Unit. 35

HFQ Frequency Entropy. xi, xiii, 23, 24, 26, 30, 60, 70, 71, 73, 76

HHH Combined Entropy. 24, 26, 60, 70, 71, 73, 76

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

HPC High-Performance Computing. vii, ix, 3, 8, 31, 49, 50, 57

HTP Temporal Entropy. xi, xiii, 24, 26, 29, 30, 60, 70, 71, 73, 76

IFFT Inverse Fast Fourier Transform. 16

IHEPC Individual Household Electric Power Consumption. 37

IQR Interquartile Range. 62–64, 66, 67

JKU Johannes Kepler Universität Linz. 36

JSON Java Script Object Notation. 49, 53–55, 60

KG Knowledge Graph. 79

KUWI Kurzfristige Wissenschaftliche Arbeiten. vii, ix

LSTM Long Short-Term Memory. xi, xiii, 5, 8, 31, 32, 35–38, 40–49, 52, 53, 55–57, 63–65, 67, 68, 72, 76, 77, 90, 93, 101, 102

MAAD Multiresolution Analysis of Acoustic Diversity. 12, 17, 79

MAX Maximum. xi, xiii, 38–40, 44, 55, 60, 62, 65–70

MIR Music Information Retrieval. 3, 5, 78, 90

MIT Massachusetts Institute of Technology. 35

ML Machine Learning. 1, 3, 5–7, 31–33, 56, 57, 90

MMM Temporal Median. 21, 22, 26, 60, 70, 73, 76

NDSI Normalized Difference Soundscape Index. 25, 26, 76

NFWA Normalised Aggregated Frequency-Weighted Accuracy. 70–73

NLP Natural Language Processing. 36

NN Neural Network. 31–34, 52, 100

NPP Number of Peaks. xi, xiii, 22, 23, 26, 29, 30, 60, 70–73, 76, 97

PAM Passive Acoustic Monitoring. xi, xiii, 1, 2, 7, 12, 78, 79

ParaNet-CNN-LSTM Parallel Network CNN - LSTM. xi, xiii, 8, 31, 37, 45, 46, 48, 53, 55–57, 63–71, 76, 77, 99

PLCNet Parallel LSTM - CNN Network. 45, 46

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

PSA Payment for Environmental Services. xi, xiii, 2, 3, 12, 36, 78

Q Quartile. 66, 77

REAL Remote Environmental Assessment Laboratory. 79

Reference Forest, Pasture, Natural Regeneration, and Plantation Regions. xi, xiii, 3, 4, 7, 12, 32, 45, 48, 76, 77

Regions Reference Forest, Pasture, Natural Regeneration, and Plantation. 12, 29, 30, 76

ReLU Rectified Linear Unit. 33, 38–40, 43–45

RNN Recurrent Neural Network. 3, 5, 34–36, 40, 44, 90

RQ Research Question. 1, 3–5, 75, 76, 84, 89–91

SCITAS Scientific IT & Application Support. vii, ix, 31, 50

SCRUM SCRUM Agile for Software Development. 3, 4, 6, 7, 82–87

SDSC Swiss Data Science Center. vii, ix

SEMP Swiss-European Mobility Programme. vii, ix

SeqNet-CNN-LSTM Sequential Network CNN - LSTM. xi, xiii, 8, 31, 37, 38, 40–45, 53, 55–58, 63–70, 76, 99

SeqNet-LSTM-CNN Sequential Network LSTM - CNN. xi, xiii, 8, 31, 37, 42–44, 47, 53, 55–58, 63–70, 76, 77, 99

SGD Stochastic Gradient Descent. 33

SLR Systematic Literature Review. xvi, 3–7, 81–87, 89–91

SOTA State of the art. 6, 8, 9

STFT Short-Time Fourier Transform. 13–15, 18

STLF Short-Term Load Forecasting. 45, 46

SUM Summation. 60

SVM Support Vector Machine. xi, xiii, 31, 48, 49, 53, 55–57, 61, 63–65, 67, 68, 77, 93

TN True Negatives. 54

TP True Positives. 54

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

TU Wien Technische Universität Wien. vii, ix, 31, 50

TUM Technische Universität München. 35

UVCGAN UNet Vision Transformer Cycle-Consistent GAN for Unpaired Image-to-Image Translation. 79

VSC Vienna Scientific Cluster. vii, ix, 31

xLSTM Extended Long Short-Term Memory. 36

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Bibliography

[1] United Nations, *The sustainable development goals report 2023*, Accessed: 2025-04-13, 2023. [Online]. Available: https://unstats.un.org/sdgs/report/2023/The-Sustainable-Development-Goals-Report-2023.pdf.

[2] M. R. W. Rands, W. M. Adams, L. Bennun, *et al.*, "Biodiversity conservation: Challenges beyond 2010", *Science*, vol. 329, no. 5997, pp. 1298–1303, 2010. DOI: 10.1126/science.1189138. eprint: https://www.science.org/doi/pdf/10.1126/science.1189138. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1189138.

[3] S. Alibakhshi, S. C. Cook-Patton, E. Davin, *et al.*, "Natural forest regeneration is projected to reduce local temperatures", *Communications Earth & Environment*, vol. 5, no. 1, p. 577, 2024, ISSN: 2662-4435. DOI: 10.1038/s43247-024-01737-5. [Online]. Available: https://doi.org/10.1038/s43247-024-01737-5.

[4] Z. Burivalova, Purnomo, B. Wahyudi, *et al.*, "Using soundscapes to investigate homogenization of tropical forest diversity in selectively logged forests", *Journal of Applied Ecology*, vol. 56, no. 11, pp. 2493–2504, 2019. DOI: https://doi.org/10.1111/1365-2664.13481. eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/1365-2664.13481. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/1365-2664.13481.

[5] J. Millward-Hopkins and Y. Oswald, "Reducing global inequality to secure human wellbeing and climate safety: A modelling study", *The Lancet Planetary Health*, vol. 7, no. 2, e147–e154, 2023, ISSN: 2542-5196. DOI: 10.1016/S2542-5196(23)00004-9. [Online]. Available: https://doi.org/10.1016/S2542-5196(23)00004-9.

[6] G. L. Delgado, J. van den Hoogen, D. H. Dent, *et al.*, "Large-scale recovery in costa rica's payment for ecosystem service program", *bioRxiv*, 2024. DOI: 10.1101/2024.09.03.610944. [Online]. Available: http://biorxiv.org/lookup/doi/10.1101/2024.09.03.610944.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[7] Convention on Biological Diversity, *Decision 14/34: Comprehensive and participatory process for the preparation of the post-2020 global biodiversity framework*, Conference of the Parties to the Convention on Biological Diversity (CBD COP-14), Sharm El-Sheikh, Egypt, Accessed: 25022025, 2018. [Online]. Available: https://www.cbd.int/doc/decisions/cop-14/cop-14-dec-34-en.pdf.

[8] D. Wuepper, T. Crowther, T. Lauber, *et al.*, "Public policies and global forest conservation: Empirical evidence from national borders", *Global Environmental Change*, vol. 84, p. 102 770, 2024, ISSN: 0959-3780. DOI: 10.1016/j.gloenvcha.2023.102770. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095937802300136X.

[9] G. Mikkelson, A. Gonzalez, and G. Peterson, "Economic inequality predicts biodiversity loss", *PloS one*, vol. 2, e444, Feb. 2007. DOI: 10.1371/journal.pone.0000444.

[10] M. Miyamoto, "Poverty reduction saves forests sustainably: Lessons for deforestation policies", *World Development*, vol. 127, p. 104 746, 2020, ISSN: 0305-750X. DOI: https://doi.org/10.1016/j.worlddev.2019.104746. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0305750X1930395X.

[11] A. Castano Garcia, A. Ambrose, A. Hawkins, and S. Parkes, "High consumption, an unsustainable habit that needs more attention", *Energy Research Social Science*, vol. 80, p. 102 241, 2021, ISSN: 2214-6296. DOI: https://doi.org/10.1016/j.erss.2021.102241. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214629621003340.

[12] P. Furumo and T. M. Aide, "Using soundscapes to assess biodiversity in neotropical oil palm landscapes", *Landscape Ecology*, vol. 34, Apr. 2019. DOI: 10.1007/s10980-019-00815-w.

[13] T. Bradfer-Lawrence, N. Gardner, L. Bunnefeld, N. Bunnefeld, S. G. Willis, and D. H. Dent, "Guidelines for the use of acoustic indices in environmental research", *Methods in Ecology and Evolution*, vol. 10, no. 10, pp. 1796–1807, 2019. DOI: https://doi.org/10.1111/2041-210X.13254. eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13254. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13254.

[14] I. Havinga, L. Hein, M. Vega-Araya, and A. Languillaume, "Spatial quantification to examine the effectiveness of payments for ecosystem services: A case study of costa rica's pago de servicios ambientales", *Ecological Indicators*, vol. 108, p. 105 766, 2020, ISSN: 1470-160X. DOI: https://doi.org/10.1016/j.ecolind.2019.105766. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1470160X19307605.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[15] L. Chancel, P. Bothe, and T. Voituriez, "The potential of wealth taxation to address the triple climate inequality crisis", *Nature Climate Change*, vol. 14, no. 1, pp. 5–7, 2024, ISSN: 1758-6798. DOI: 10.1038/s41558-023-01891-2. [Online]. Available: https://doi.org/10.1038/s41558-023-01891-2.

[16] T. Bradfer-Lawrence, T. Finch, R. B. Bradbury, G. M. Buchanan, A. Midgley, and R. H. Field, "The potential contribution of terrestrial nature-based solutions to a national "net zero" climate target", *Journal of Applied Ecology*, vol. 58, no. 11, pp. 2349–2360, 2021. DOI: 10.1111/1365-2664.14003. eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/1365-2664.14003. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/1365-2664.14003.

[17] O. Sánchez-Chaves and G. Navarrete-Chacón, "La experiencia de costa rica en el pago por servicios ambientales: 20 años de lecciones aprendidas", *Revista de Ciencias Ambientales*, vol. 51, no. 2, pp. 195–214, 2017. DOI: 10.15359/rca.51-2.11. [Online]. Available: https://www.revistas.una.ac.cr/index.php/ambientales/article/view/9487.

[18] R. Fletcher and J. Breitling, "Market mechanism or subsidy in disguise? governing payment for environmental services in costa rica", *Geoforum*, vol. 43, no. 3, pp. 402–411, 2012, The Global Rise and Local Implications of Market-Oriented Conservation Governance, ISSN: 0016-7185. DOI: 10.1016/j.geoforum.2011.11.008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0016718511002107.

[19] N. van Tiel, F. Fopp, P. Brun, *et al.*, "Regional uniqueness of tree species composition and response to forest loss and climate change", *Nature Communications*, vol. 15, no. 1, p. 4375, 2024, ISSN: 2041-1723. DOI: 10.1038/s41467-024-48276-3. [Online]. Available: https://doi.org/10.1038/s41467-024-48276-3.

[20] J. Sueur and A. Farina, "Ecoacoustics: The ecological investigation and interpretation of environmental sound", *Biosemiotics*, vol. 8, no. 3, pp. 493–502, 2015. DOI: 10.1007/s12304-015-9248-x. [Online]. Available: https://doi.org/10.1007/s12304-015-9248-x.

[21] T. Luypaert, A. S. Bueno, G. S. Masseli, *et al.*, "A framework for quantifying soundscape diversity using hill numbers", *Methods in Ecology and Evolution*, vol. 13, no. 10, pp. 2262–2274, 2022. DOI: https://doi.org/10.1111/2041-210X.13924. eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13924. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13924.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[22] Z. Burivalova, T. Maeda, Purnomo, *et al.*, "Loss of temporal structure of tropical soundscapes with intensifying land use in borneo", *Science of The Total Environment*, vol. 852, p. 158 268, 2022, ISSN: 0048-9697. DOI: https://doi.org/10.1016/j.scitotenv.2022.158268. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0048969722053670.

[23] Z. Burivalova, M. Towsey, T. Boucher, *et al.*, "Using soundscapes to detect variable degrees of human influence on tropical forests in papua new guinea", *Conservation Biology*, vol. 32, pp. 205–215, 1 Feb. 2018, ISSN: 15231739. DOI: 10.1111/cobi.12968.

[24] V. Ramesh, P. Hariharan, V. Akshay, *et al.*, "Using passive acoustic monitoring to examine the impacts of ecological restoration on faunal biodiversity in the western ghats", *Biological Conservation*, vol. 282, p. 110 071, 2023, ISSN: 0006-3207. DOI: https://doi.org/10.1016/j.biocon.2023.110071. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0006320723001726.

[25] P. Roe, P. Eichinski, R. A. Fuller, *et al.*, "The australian acoustic observatory", *Methods in Ecology and Evolution*, vol. 12, no. 10, pp. 1802–1808, 2021. DOI: https://doi.org/10.1111/2041-210X.13660. eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13660. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13660.

[26] J. Müller, O. Mitesser, H. M. Schaefer, *et al.*, "Soundscapes and deep learning enable tracking biodiversity recovery in tropical forests", *Nature Communications*, vol. 14, no. 1, p. 6191, 2023, ISSN: 2041-1723. DOI: 10.1038/s41467-023-41693-w. [Online]. Available: https://doi.org/10.1038/s41467-023-41693-w.

[27] M. W. Towsey, "The calculation of acoustic indices derived from long-duration recordings of the natural environment", 2017. [Online]. Available: https://eprints.qut.edu.au/110634/.

[28] R. Gibb, E. Browning, P. Glover-Kapfer, and K. E. Jones, "Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring", *Methods in Ecology and Evolution*, vol. 10, no. 2, pp. 169–185, 2019. DOI: https://doi.org/10.1111/2041-210X.13101. eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13101. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13101.

[29] A. Taherkhani, G. Cosma, and T. M. McGinnity, "A deep convolutional neural network for time series classification with intermediate targets", *SN Computer Science*, vol. 4, no. 6, p. 832, 2023, ISSN: 2661-8907. DOI: 10.1007/s42979-023-02159-4. [Online]. Available: https://doi.org/10.1007/s42979-023-02159-4.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[30] J. Poutaraud, J. Sueur, C. Thébaud, and S. Haupert, "Meta-embedded clustering (mec): A new method for improving clustering quality in unlabeled bird sound datasets", *Ecological Informatics*, vol. 82, p. 102 687, Jun. 2024. DOI: 10.1016/j.ecoinf.2024.102687.

[31] M. A. R. Ken Peffers Tuure Tuunanen and S. Chatterjee, "A design science research methodology for information systems research", *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, 2007. DOI: 10.2753/MIS0742-1222240302. eprint: https://doi.org/10.2753/MIS0742-1222240302. [Online]. Available: https://doi.org/10.2753/MIS0742-1222240302.

[32] E. Engström, M.-A. Storey, P. Runeson, M. Höst, and M. T. Baldassarre, "How software engineering research aligns with design science: A review", *Empirical Software Engineering*, vol. 25, no. 4, pp. 2630–2660, 2020, ISSN: 1573-7616. DOI: 10.1007/s10664-020-09818-7. [Online]. Available: https://doi.org/10.1007/s10664-020-09818-7.

[33] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research", *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004, ISSN: 02767783. [Online]. Available: http://www.jstor.org/stable/25148625 (visited on 02/09/2025).

[34] H. Snyder, "Literature review as a research methodology: An overview and guidelines", *Journal of Business Research*, vol. 104, pp. 333–339, 2019, ISSN: 0148-2963. DOI: https://doi.org/10.1016/j.jbusres.2019.07.039. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0148296319304564.

[35] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying crisp-dm process model", *Procedia Computer Science*, vol. 181, pp. 526–534, 2021, CENTERIS 2020 - International Conference on ENTERprise Information Systems / ProjMAN 2020 - International Conference on Project MANagement / HCist 2020 - International Conference on Health and Social Care Information Systems and Technologies 2020, CENTERIS/ProjMAN/HCist 2020, ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2021.01.199. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050921002416.

[36] B. A. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering", English, Keele University and Durham University Joint Report, Tech. Rep. EBSE 2007-001, 2007. DOI: 10.1145/2372233.2372235. [Online]. Available: https://doi.org/10.1145/2372233.2372235.

[37] M.-S. Stupariu, S. A. Cushman, A.-I. Plesoianu, I. Patru-Stupariu, and C. Furst, "Machine learning in landscape ecological analysis: A review of recent approaches", *Landscape Ecology*, vol. 37, no. 5, pp. 1227–1250, 2022, ISSN: 1572-9761. DOI:

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

10.1007/s10980-021-01366-9. [Online]. Available: https://doi.org/10.1007/s10980-021-01366-9.

[38] A. Di Vaio, R. Palladino, R. Hassan, and O. Escobar, "Artificial intelligence and business models in the sustainable development goals perspective: A systematic literature review", *Journal of Business Research*, vol. 121, pp. 283–314, 2020, ISSN: 0148-2963. DOI: 10.1016/j.jbusres.2020.08.019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0148296320305191.

[39] P. Chapman, "Crisp-dm 1.0: Step-by-step data mining guide", 2000. [Online]. Available: https://api.semanticscholar.org/CorpusID:59777418.

[40] R. Wirth and J. Hipp, "Crisp-dm: Towards a standard process model for data mining", in *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 2000. [Online]. Available: https://api.semanticscholar.org/CorpusID:1211505.

[41] PACKT Publishing, *Comprehensive guide to agile scrum foundations*, eng, Online course, Accessed: 2025-04-13, 2024.

[42] K. Schwaber, "Scrum development process", in *Business Object Design and Implementation*, J. Sutherland, C. Casanave, J. Miller, P. Patel, and G. Hollowell, Eds., London: Springer London, 1997, pp. 117–134, ISBN: 978-1-4471-0947-1. DOI: 10.1007/978-1-4471-0947-1_11. [Online]. Available: https://doi.org/10.1007/978-1-4471-0947-1_11.

[43] J. S. Ulloa, S. Haupert, J. F. Latorre, T. Aubin, and J. Sueur, "Scikit-maad: An open-source and modular toolbox for quantitative soundscape analysis in python", *Methods in Ecology and Evolution*, vol. 12, no. 12, pp. 2334–2340, 2021. DOI: 10.1111/2041-210X.13711. eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13711. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13711.

[44] A. Paszke, S. Gross, F. Massa, *et al.*, "Pytorch: An imperative style, high-performance deep learning library", *CoRR*, vol. abs/1912.01703, 2019. arXiv: 1912.01703. [Online]. Available: http://arxiv.org/abs/1912.01703.

[45] T. Sainburg, M. Thielk, and T. Q. Gentner, "Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires", *PLoS Computational Biology*, vol. 16, no. 10, e1008228, 2020.

[46] T. Sainburg, *Timsainb/noisereduce: V1.0*, version db94fe2, 2019. DOI: 10.5281/zenodo.3243139. [Online]. Available: https://doi.org/10.5281/zenodo.3243139.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[47] N. Yoh, C. L. Haley, and Z. Burivalova, "Time series methods for the analysis of soundscapes and other cyclical ecological data", *Methods in Ecology and Evolution*, vol. 15, no. 7, pp. 1158–1176, 2024. DOI: 10.1111/2041-210X.14361. eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.14361. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.14361.

[48] P. Kahlig, "Some aspects of julius von hann's contribution to modern climatology", in *Interactions Between Global Climate Subsystems*. American Geophysical Union (AGU), 1993, pp. 1–7, ISBN: 9781118666593. DOI: 10.1029/GM075p0001. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/GM075p0001. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/GM075p0001.

[49] R. B. Blackman and J. W. Tukey, "The measurement of power spectra from the point of view of communications engineering — part i", *The Bell System Technical Journal*, vol. 37, no. 1, pp. 185–282, 1958. DOI: 10.1002/j.1538-7305.1958.tb03874.x.

[50] N. Pielawski and C. Wählby, "Introducing hann windows for reducing edge-effects in patch-based image segmentation", *PLOS ONE*, vol. 15, no. 3, J. Zhang, Ed., e0229839, 2020, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0229839. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0229839.

[51] A. Arranz-Gimon, A. Zorita-Lamadrid, D. Morinigo-Sotelo, and O. Duque-Perez, "Analysis of the use of the hanning window for the measurement of interharmonic distortion caused by close tones in iec standard framework", *Electric Power Systems Research*, vol. 206, p. 107 833, 2022, ISSN: 0378-7796. DOI: 10.1016/j.epsr.2022.107833. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378779622000633.

[52] A. A. Alkahtani, S. T. Y. Alfalahi, A. A. Athamneh, *et al.*, "Power quality in microgrids including supraharmonics: Issues, standards, and mitigations", *IEEE Access*, vol. 8, pp. 127 104–127 122, 2020. DOI: 10.1109/ACCESS.2020.3008042.

[53] T. Bradfer-Lawrence, N. Gardner, L. Bunnefeld, N. Bunnefeld, S. G. Willis, and D. H. Dent, "Guidelines for the use of acoustic indices in environmental research", *Methods in Ecology and Evolution*, vol. 10, no. 10, pp. 1796–1807, 2019. DOI: 10.1111/2041-210X.13254. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13254.

[54] T. Bradfer-Lawrence, C. Desjonqueres, A. Eldridge, A. Johnston, and O. Metcalf, "Using acoustic indices in ecology: Guidance on study design, analyses and interpretation", *Methods in Ecology and Evolution*, vol. 14, no. 9, pp. 2192–2204, 2023. DOI: 10.1111/2041-210X.14194. eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.14194. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.14194.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[55] T. Bradfer-Lawrence, B. Duthie, C. Abrahams, *et al.*, "The acoustic index user's guide: A practical manual for defining, generating and understanding current and future acoustic indices", *Methods in Ecology and Evolution*, vol. n/a, no. n/a, 2023. DOI: 10.1111/2041-210X.14357. eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.14357. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.14357.

[56] L. J. Villanueva-Rivera, B. C. Pijanowski, J. Doucette, and B. Pekin, "A primer of acoustic analysis for landscape ecologists", *Landscape Ecology*, vol. 26, no. 9, pp. 1233–1246, 2011, ISSN: 1572-9761. DOI: 10.1007/s10980-011-9636-9. [Online]. Available: https://doi.org/10.1007/s10980-011-9636-9.

[57] B. C. Pijanowski, L. J. Villanueva-Rivera, S. L. Dumyahn, *et al.*, "Soundscape ecology: The science of sound in the landscape", *BioScience*, vol. 61, no. 3, pp. 203–216, 2011, ISSN: 0006-3568, 1525-3244. DOI: 10.1525/bio.2011.61.3.6. [Online]. Available: https://doi.org/10.1525/bio.2011.61.3.6 (visited on 04/18/2025).

[58] D. Luna-Naranjo, J. D. Martinez-Vargas, C. Sánchez-Giraldo, J. M. Daza, and J. D. López, "Quantifying and mitigating recorder-induced variability in ecological acoustic indices", *Ecological Informatics*, vol. 82, p. 102 668, 2024, ISSN: 1574-9541. DOI: 10.1016/j.ecoinf.2024.102668. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574954124002103.

[59] I. Alcocer, H. Lima, L. S. M. Sugai, and D. Llusia, "Acoustic indices as proxi+es for biodiversity: A meta-analysis", *Biological Reviews*, vol. 97, no. 6, pp. 2209–2236, 2022. DOI: 10.1111/brv.12890. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/brv.12890. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/brv.12890.

[60] M. Depraetere, S. Pavoine, F. Jiguet, A. Gasc, S. Duvail, and J. Sueur, "Monitoring animal diversity using acoustic indices: Implementation in a temperate woodland", *Ecological Indicators*, vol. 13, no. 1, pp. 46–54, 2012, ISSN: 1470-160X. DOI: 10.1016/j.ecolind.2011.05.006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1470160X11001361.

[61] N. Pieretti, A. Farina, and D. Morri, "A new methodology to infer the singing activity of an avian community: The acoustic complexity index (aci)", *Ecological Indicators*, vol. 11, no. 3, pp. 868–873, 2011, ISSN: 1470-160X. DOI: 10.1016/j.ecolind.2010.11.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1470160X10002037.

[62] A. Farina, N. Pieretti, P. Salutari, E. Tognari, and A. Lombardi, "The application of the acoustic complexity indices (aci) to ecoacoustic event detection and identification (eedi) modeling", *Biosemiotics*, vol. 9, pp. 227–246, 2 Aug. 2016, ISSN: 18751350. DOI: 10.1007/s12304-016-9266-3.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[63] N. T. Boelman, G. P. Asner, P. J. Hart, and R. E. Martin, "Multi-trophic invasion resistance in hawaii: Bioacoustics, field surveys, and airborne remote sensing", *Ecological Applications*, vol. 17, no. 8, pp. 2137–2144, 2007. DOI: 10.1890/07-0004.1. eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1890/07-0004.1. [Online]. Available: https://esajournals.onlinelibrary.wiley.com/doi/10.1890/07-0004.1.

[64] M. Johansson, "The hilbert transform", Master's Thesis, M.S. thesis, Växjö University, 1999. [Online]. Available: http://w3.msi.vxu.se/exarb/mj_ex.pdf.

[65] A. Gasc, J. Sueur, S. Pavoine, R. Pellens, and P. Grandcolas, "Biodiversity sampling using a global acoustic approach: Contrasting sites with microendemics in new caledonia", *PLOS ONE*, vol. 8, no. 5, pp. 1–10, 2013. DOI: 10.1371/journal.pone.0065311. [Online]. Available: https://doi.org/10.1371/journal.pone.0065311.

[66] J. Sueur, T. Aubin, and C. Simonis, "Seewave: A free modular tool for sound analysis and synthesis", *Bioacoustics*, vol. 18, no. 2, pp. 213–226, 2008. DOI: 10.1080/09524622.2008.9753600. eprint: https://doi.org/10.1080/09524622.2008.9753600. [Online]. Available: https://doi.org/10.1080/09524622.2008.9753600.

[67] J. Sueur, S. Pavoine, O. Hamerlynck, and S. Duvail, "Rapid acoustic survey for biodiversity appraisal", *PLOS ONE*, vol. 3, no. 12, pp. 1–9, 2009. DOI: 10.1371/journal.pone.0004065. [Online]. Available: https://doi.org/10.1371/journal.pone.0004065.

[68] C. E. Shannon, "A mathematical theory of communication", *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[69] C. Gini, *Variabilità e mutabilità*, Reprinted in *Memorie di Metodologica Statistica*, ed. E. Pizetti and T. Salvemini, Rome: Libreria Eredi Virgilio Veschi, 1912.

[70] E. P. Kasten, S. H. Gage, J. Fox, and W. Joo, "The remote environmental assessment laboratory's acoustic library: An archive for studying soundscape ecology", *Ecological Informatics*, vol. 12, pp. 50–67, 2012, ISSN: 1574-9541. DOI: 10.1016/j.ecoinf.2012.08.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S157495411200088X.

[71] EPFL SCITAS, *Izar – epfl supercomputer*, Accessed: 2025-07-03, 2024. [Online]. Available: https://www.epfl.ch/research/facilities/scitas/hardware/izar/.

[72] Vienna Scientific Cluster, *Vsc-5 – austria's most powerful supercomputer*, Accessed: 2025-07-03, 2022. [Online]. Available: https://www.vsc.ac.at/systems/vsc-5/.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[73] K. P. Murphy, *Probabilistic Machine Learning: An Introduction* (Adaptive Computation and Machine Learning). Cambridge, Massachusetts: The MIT Press, 2022, ISBN: 0262046822.

[74] T. M. Mitchell, *Machine Learning* (McGraw-Hill Series in Computer Science: Artificial Intelligence). New York, NY: McGraw-Hill, 1997, ISBN: 0071154671.

[75] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (Adaptive Computation and Machine Learning). Cambridge, MA; London, England: The MIT Press, 2016, ISBN: 0262035618.

[76] E. Charniak, *Introduction to Deep Learning*. Cambridge, MA; London, England: The MIT Press, 2018, ISBN: 0262039516.

[77] C. C. Aggarwal, *Neural Networks and Deep Learning: A Textbook*, eng, Second edition. Cham: Springer, 2023, ISBN: 3031296419.

[78] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel: Rapid training data creation with weak supervision", *Proceedings of the VLDB Endowment*, vol. 11, no. 3, pp. 269–282, 2017, ISSN: 2150-8097. DOI: 10.14778/3157794.3157797. [Online]. Available: https://doi.org/10.14778/3157794.3157797.

[79] H. Chen, J. Wang, L. Feng, *et al.*, "A general framework for learning from weak supervision", *arXiv preprint arXiv:2402.01922*, 2024. [Online]. Available: https://arxiv.org/abs/2402.01922.

[80] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 79, no. 8, pp. 2554–2558, 1982. [Online]. Available: http://www.jstor.org/stable/12175.

[81] G. E. Hinton, *Connectionist Symbol Processing* (Special Issues of Artificial Intelligence), en, 1st MIT Press ed. Cambridge, MA: MIT Press, 1991, ISBN: 0262256363. [Online]. Available: https://ieeexplore.ieee.org/servlet/opac?bknumber=6267281.

[82] G. Hinton, *Unsupervised Learning: Foundations of Neural Computation* (Computational Neuroscience), en. Cambridge, MA: MIT Press, 1999, ISBN: 026258168X.

[83] V. Lakshmanan, S. Robinson, and M. Munn, *Machine Learning Design Patterns: Solutions to Common Challenges in Data Preparation, Model Building, and MLOps*, eng, 1st edition. Beijing: O'Reilly, 2021, ISBN: 1098115732.

[84] C. M. Bishop, *Neural Networks for Pattern Recognition*, en, 1st ed., reprint. Oxford: Clarendon Press, 1997, ISBN: 0198538642. DOI: 10.1145/294828.1067910.

[85] Royal Swedish Academy of Sciences, *The nobel prize in physics 2024: They trained artificial neural networks using physics*, Accessed: 2025-08-21, KVA, 2024. [Online]. Available: https://www.kva.se/en/news/the-nobel-prize-in-physics-2024/.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[86] B. Li and S. Gilbert, "Artificial intelligence awarded two nobel prizes for innovations that will shape the future of medicine", eng, *npj Digital Medicine*, vol. 7, no. 1, p. 336, 2024, ISSN: 2398-6352. DOI: 10.1038/s41746-024-01345-9. [Online]. Available: https://doi.org/10.1038/s41746-024-01345-9.

[87] S. Ruder, *An overview of gradient descent optimization algorithms*, 2017. arXiv: 1609.04747 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1609.04747.

[88] Y. LeCun, B. Boser, J. S. Denker, *et al.*, "Backpropagation applied to handwritten zip code recognition", *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989, ISSN: 0899-7667.

[89] C. M. Bishop, *Pattern Recognition and Machine Learning* (Information Science and Statistics). New York, NY: Springer, 2006, ISBN: 0387310738. DOI: 10.1117/1.2819119.

[90] D. Scherer, A. Müller, and S. Behnke, "Evaluation of pooling operations in convolutional architectures for object recognition", in *Artificial Neural Networks – ICANN 2010*, K. Diamantaras, W. Duch, and L. S. Iliadis, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 92–101, ISBN: 978-3-642-15825-4. DOI: 10.1007/978-3-642-15825-4_10.

[91] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. N. Sainath, "Deep learning for audio signal processing", *CoRR*, vol. abs/1905.00078, 2019. arXiv: 1905.00078 [cs.SD]. [Online]. Available: http://arxiv.org/abs/1905.00078.

[92] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting", eng, *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014, ISSN: 1532-4435.

[93] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017, ISSN: 0001-0782. DOI: 10.1145/3065386. [Online]. Available: https://doi.org/10.1145/3065386.

[94] A. Rius, I. Ruisánchez, M. P. Callao, and F. X. Rius, "Reliability of analytical systems: Use of control charts, time series models and recurrent neural networks (rnn)", *Chemometrics and Intelligent Laboratory Systems*, vol. 40, no. 1, pp. 1–18, 1998. DOI: 10.1016/S0169-7439(97)00085-3. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169743997000853.

[95] J. Hou, Y. Wang, J. Zhou, and Q. Tian, "Prediction of hourly air temperature based on cnn-lstm", *Geomatics, Natural Hazards and Risk*, vol. 13, no. 1, pp. 1962–1986, 2022, ISSN: 1947-5705. DOI: 10.1080/19475705.2022.2102942.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

[96] L. Nie, Y. Zhang, and H. Wang, "Classification of underwater soundscapes using raw hydroacoustic signals", *The Journal of the Acoustical Society of America*, vol. 154, no. 4$_s$*upplement*, A304–A304, Oct. 2023, ISSN: 0001-4966. DOI: 10.1121/10.0023608. [Online]. Available: https://doi.org/10.1121/10.0023608.

[97] X. Lin, R. Dong, and Z. Lv, "Deep learning-based classification of raw hydroacoustic signal: A review", *Journal of Marine Science and Engineering*, vol. 11, no. 1, 2023, ISSN: 2077-1312. DOI: 10.3390/jmse11010003. [Online]. Available: https://www.mdpi.com/2077-1312/11/1/3.

[98] S. Lapp, T. Rhinehart, L. Freeland-Haynes, J. Khilnani, A. Syunkova, and J. Kitzes, "Opensoundscape: An open-source bioacoustics analysis package for python", *Methods in Ecology and Evolution*, vol. 14, no. 9, pp. 2321–2328, 2023. DOI: 10.1111/2041-210X.14196. eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.14196. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.14196.

[99] S. Cominelli, N. Bellin, C. D. Brown, V. Rossi, and J. Lawson, "Acoustic features as a tool to visualize and explore marine soundscapes: Applications illustrated using marine mammal passive acoustic monitoring datasets", *Ecology and Evolution*, vol. 14, no. 2, e10951, 2024, e10951 ECE-2022-08-01269.R2. DOI: 10.1002/ece3.10951. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ece3.10951. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.10951.

[100] J. Shi, S. Wang, P. Qu, and J. Shao, "Time series prediction model using lstm-transformer neural network for mine water inflow", *Scientific Reports*, vol. 14, no. 1, p. 18 284, 2024, ISSN: 2045-2322. DOI: 10.1038/s41598-024-69418-z.

[101] G. Li, Y. Li, G. Han, *et al.*, "Forecasting and analyzing influenza activity in hebei province, china, using a cnn-lstm hybrid model", *BMC Public Health*, vol. 24, no. 1, p. 2171, 2024. DOI: 10.1186/s12889-024-19590-8. [Online]. Available: https://doi.org/10.1186/s12889-024-19590-8.

[102] C. A. Quinn, P. Burns, G. Gill, *et al.*, "Soundscape classification with convolutional neural networks reveals temporal and geographic patterns in ecoacoustic data", *Ecological Indicators*, vol. 138, p. 108 831, 2022, ISSN: 1470-160X. DOI: https://doi.org/10.1016/j.ecolind.2022.108831. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1470160X22003028.

[103] *Recurrent Neural Networks for Prediction*. John Wiley & Sons, Ltd, 2001, ISBN: 9780470845356. DOI: 10.1002/047084535X. [Online]. Available: https://onlinelibrary.wiley.com/doi/book/10.1002/047084535X.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[104] A. Graves, "Long short-term memory", in *Supervised Sequence Labelling with Recurrent Neural Networks*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 37–45, ISBN: 978-3-642-24797-2. DOI: 10.1007/978-3-642-24797-2_4. [Online]. Available: https://doi.org/10.1007/978-3-642-24797-2_4.

[105] H. Salehinejad, S. Sankar, J. Barfett, E. Colak, and S. Valaee, "Recent advances in recurrent neural networks", *arXiv preprint*, vol. arXiv:1801.01078, 2018. arXiv: 1801.01078 [cs.NE]. [Online]. Available: https://arxiv.org/abs/1801.01078.

[106] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, *Learning phrase representations using rnn encoder-decoder for statistical machine translation*, 2014. arXiv: 1406.1078 [cs.CL]. [Online]. Available: http://arxiv.org/abs/1406.1078.

[107] S. Hochreiter and J. Schmidhuber, "Long short-term memory", *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997, ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. eprint: https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735.

[108] H. Phan, P. Koch, F. Katzberg, M. Maass, R. Mazur, and A. Mertins, *Audio scene classification with deep recurrent neural networks*, 2017. arXiv: 1703.04770 [cs.SD]. [Online]. Available: https://arxiv.org/abs/1703.04770.

[109] S. Hochreiter, "Untersuchungen zu dynamischen neuronalen netzen", Master's Thesis, Technische Universität München, 1991.

[110] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition", *arXiv preprint arXiv:1402.1128*, 2014. [Online]. Available: https://arxiv.org/abs/1402.1128.

[111] B. Dong, C. Lumezanu, Y. Chen, *et al.*, "At the speed of sound: Efficient audio scene classification", in *Proceedings of the 2020 International Conference on Multimedia Retrieval*, ser. ICMR '20, Dublin, Ireland: Association for Computing Machinery, 2020, pp. 301–305, ISBN: 9781450370875. DOI: 10.1145/3372278.3390730. [Online]. Available: https://doi.org/10.1145/3372278.3390730.

[112] M. Beck, K. Pöppel, M. Spanring, *et al.*, "Xlstm: Extended long short-term memory", *arXiv preprint arXiv:2405.04517*, 2024. arXiv: 2405.04517 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2405.04517.

[113] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need", *arXiv preprint arXiv:1706.03762*, 2017. arXiv: 1706.03762 [cs.CL]. [Online]. Available: http://arxiv.org/abs/1706.03762.

[114] C. M. Bishop and H. Bishop, *Deep Learning: Foundations and Concepts* (Information Science and Statistics), 1st ed. Cham: Springer International Publishing, 2024, ISBN: 9783031454677. DOI: 10.1007/978-3-031-45468-4. [Online]. Available: https://doi.org/10.1007/978-3-031-45468-4.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[115] Z. Wang, Y. Li, L. Zan, Z. Gong, and M. Zhu, *Stoxlstm: A stochastic extended long short-term memory network for time series forecasting*, 2025. arXiv: 2509.01187 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2509.01187.

[116] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks", eng, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2015-, IEEE, 2015, pp. 559–563, ISBN: 1467369977.

[117] S. Hershey, S. Chaudhuri, D. P. W. Ellis, *et al.*, *Cnn architectures for large-scale audio classification*, 2017. arXiv: 1609.09430 [cs.SD]. [Online]. Available: https://arxiv.org/abs/1609.09430.

[118] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection", *Applied Sciences*, vol. 6, no. 6, pp. 162–162, 2016, ISSN: 2076-3417.

[119] A. Mesaros, A. Diment, B. Elizalde, *et al.*, "Sound event detection in the dcase 2017 challenge", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 992–1006, 2019, ISSN: 2329-9290. DOI: 10.1109/TASLP.2019.2908067.

[120] A. Mesaros, T. Heittola, E. Benetos, *et al.*, "Detection and classification of acoustic scenes and events: Outcome of the dcase 2016 challenge", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, 2018, ISSN: 2329-9290.

[121] A. Mesaros, T. Heittola, and D. Ellis, "Datasets and evaluation", eng, in *Computational Analysis of Sound Scenes and Events*, Cham: Springer International Publishing, 2018, pp. 147–179, ISBN: 9783319634494.

[122] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in dcase 2019", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 684–698, 2021, ISSN: 2329-9290. DOI: 10.1109/TASLP.2020.3046174. [Online]. Available: https://doi.org/10.1109/TASLP.2020.3046174.

[123] M. Liu, K. Liang, D. Hu, *et al.*, "Tmac: Temporal multi-modal graph learning for acoustic event classification", eng, in *Proceedings of the 31st ACM International Conference on Multimedia (MM 2023)*, New York, NY, USA: ACM, 2023, pp. 3365–3374, ISBN: 9798400701085. DOI: 10.1145/3581783.3611853. [Online]. Available: https://github.com/MGitHubL/TMac.

[124] G. Tang, R. Liang, Y. Xie, Y. Bao, and S. Wang, "Improved convolutional neural networks for acoustic event classification", *Multimedia Tools and Applications*, vol. 78, no. 12, pp. 15 801–15 816, 2019, ISSN: 1573-7721. DOI: 10.1007/s11042-018-6991-4. [Online]. Available: https://doi.org/10.1007/s11042-018-6991-4.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[125] V. K. Singh, K. Sharma, and S. N. Sur, "Impact of various continuous wavelet transforms for acoustic scene classification with dcase dataset", *Signal, Image and Video Processing*, vol. 19, no. 6, p. 440, 2025, ISSN: 1863-1711. DOI: 10.1007/s11760-025-04031-9. [Online]. Available: https://doi.org/10.1007/s11760-025-04031-9.

[126] V. K. Singh, K. Sharma, and S. N. Sur, "Acoustic scene classification using dynamic time warping technique based on short time fourier transform and discrete wavelet transforms", *Circuits, Systems, and Signal Processing*, vol. 44, no. 3, pp. 1887–1913, 2025, ISSN: 1531-5878. DOI: 10.1007/s00034-024-02895-9. [Online]. Available: https://doi.org/10.1007/s00034-024-02895-9.

[127] C. Zhang, Y. Wei, X. Wang, X. Wu, and X. Zhu, "Machine sound anomaly detection based on dual-channel feature fusion variational auto-encoder", *Applied Intelligence*, vol. 55, no. 7, p. 594, 2025, ISSN: 1573-7497. DOI: 10.1007/s10489-025-06449-7. [Online]. Available: https://doi.org/10.1007/s10489-025-06449-7.

[128] H. Yang, R. Dong, R. Guo, *et al.*, "Real-time acoustic scene recognition for elderly daily routines using edge-based deep learning", *Sensors (Basel, Switzerland)*, vol. 25, no. 6, p. 1746, 2025, ISSN: 1424-8220. DOI: 10.3390/s25061746. [Online]. Available: https://doi.org/10.3390/s25061746.

[129] R. Huang, Y. Xie, and P. Jiang, "Local time-frequency feature fusion using cross-attention for acoustic scene classification", *Symmetry (Basel)*, vol. 17, no. 1, p. 49, 2025, ISSN: 2073-8994. DOI: 10.3390/sym17010049. [Online]. Available: https://doi.org/10.3390/sym17010049.

[130] A. Mesaros, R. Serizel, T. Heittola, T. Virtanen, and M. D. Plumbley, "A decade of dcase: Achievements, practices, evaluations and future challenges", eng, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Piscataway: IEEE, 2025, ISBN: 9798350368741. DOI: 10.1109/ICASSP49660.2025.10887673.

[131] L. Pham, H. Phan, R. Palaniappan, A. Mertins, and I. McLoughlin, "Cnn-moe based framework for classification of respiratory anomalies and lung disease detection", eng, *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 8, pp. 2938–2947, 2021, ISSN: 2168-2194. DOI: 10.1109/JBHI.2020.3048033.

[132] R. Akter, M. R. Islam, S. K. Debnath, P. K. Sarker, and M. K. Uddin, "A hybrid cnn-lstm model for environmental sound classification: Leveraging feature engineering and transfer learning", *Digital Signal Processing*, vol. 163, p. 105 234, 2025, ISSN: 1051-2004. DOI: 10.1016/j.dsp.2025.105234. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1051200425002568.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[133] K. M. Bellisario and B. C. Pijanowski, "Contributions of MIR to soundscape ecology. part I: Potential methodological synergies", *Ecological Informatics*, vol. 51, pp. 96–102, 2019, ISSN: 1574-9541. DOI: 10.1016/j.ecoinf.2019.02.009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574954118301912.

[134] Y. Li, T. Cai, Y. Zhang, D. Chen, and D. Dey, *What makes convolutional models great on long sequence modeling?*, 2022. arXiv: 2210.09298 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2210.09298.

[135] B. Farsi, M. Amayri, N. Bouguila, and U. Eicker, "On short-term load forecasting using machine learning techniques and a novel parallel deep lstm-cnn approach", *IEEE Access*, vol. 9, pp. 31 191–31 212, 2021, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3060290.

[136] W. Yang, Y. Huang, K. Jiang, Z. Zhang, K. Zong, and Q. Ruan, "Method of predicting braking intention using lstm-cnn-attention with hyperparameters optimized by genetic algorithm", *International Journal of Control, Automation and Systems*, vol. 22, no. 7, pp. 2301–2312, 2024. DOI: 10.1007/s12555-021-1113-x. [Online]. Available: https://doi.org/10.1007/s12555-021-1113-x.

[137] T. Le, M. T. Vo, B. Vo, E. Hwang, S. Rho, and S. W. Baik, "Improving electric energy consumption prediction using cnn and bi-lstm", *Applied Sciences*, vol. 9, no. 20, p. 4237, 2019. DOI: 10.3390/app9204237. [Online]. Available: https://www.mdpi.com/2076-3417/9/20/4237.

[138] H. Widiputra, A. Mailangkay, and E. Gautama, "Multivariate cnn-lstm model for multiple parallel financial time-series prediction", *Complexity*, vol. 2021, no. 1, p. 9 903 518, 2021. DOI: 10.1155/2021/9903518. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2021/9903518. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1155/2021/9903518.

[139] L. Zhang, Y. Cai, H. Huang, A. Li, L. Yang, and C. Zhou, "A cnn-lstm model for soil organic carbon content prediction with long time series of modis-based phenological variables", *Remote Sensing*, vol. 14, no. 18, p. 4441, 2022. DOI: 10.3390/rs14184441. [Online]. Available: https://www.mdpi.com/2072-4292/14/18/4441.

[140] M. A. B. Syed and I. Ahmed, *A cnn-lstm architecture for marine vessel track association using automatic identification system (ais) data*, 2023. arXiv: 2303.14068 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2303.14068.

[141] J. Shi, S. Wang, P. Qu, and J. Shao, "Time series prediction model using lstm-transformer neural network for mine water inflow", *Scientific Reports*, vol. 14, no. 1, p. 18 284, 2024. DOI: 10.1038/s41598-024-69418-z. [Online]. Available: https://doi.org/10.1038/s41598-024-69418-z.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[142] N. Gaud, M. Rathore, and U. Suman, "Mhcnls-har: Multiheaded cnn-lstm-based human activity recognition leveraging a novel wearable edge device for elderly health care", eng, *IEEE Sensors Journal*, vol. 24, no. 21, pp. 35 394–35 405, 2024, ISSN: 1530-437X.

[143] A. Barthwal and A. K. Goel, "Advancing air quality prediction models in urban india: A deep learning approach integrating dcnn and lstm architectures for aqi time-series classification", *Modeling Earth Systems and Environment*, vol. 10, no. 2, pp. 2935–2955, 2024, ISSN: 2363-6211. DOI: 10.1007/s40808-023-01934-9. [Online]. Available: https://doi.org/10.1007/s40808-023-01934-9.

[144] F. Yu, Y. Huang, C. Du, F. Li, Y. Gong, and G. Chen, "A cnn-lstm model for road surface recognition of electric balance vehicles", English, *Systems Science & Control Engineering*, vol. 12, no. 1, 2024, ISSN: 2164-2583. DOI: 10.1080/21642583.2023.2293912.

[145] J. Ver Berne, S. B. Saadi, N. Oliveira-Santos, L. E. Marinho-Vieira, and R. Jacobs, "Automated classification of panoramic radiographs with inflammatory periapical lesions using a cnn-lstm architecture", *Journal of Dentistry*, vol. 156, p. 105 688, 2025, ISSN: 0300-5712. DOI: 10.1016/j.jdent.2025.105688. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0300571225001332.

[146] L. Pham, D. Ngo, T. Hoang, A. Schindler, and I. McLoughlin, *An ensemble of deep learning frameworks applied for predicting respiratory anomalies*, 2022. arXiv: 2201.03054 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2201.03054.

[147] G. Hebrail and A. Berard, *Individual household electric power consumption*, UCI Machine Learning Repository, DOI: https://doi.org/10.24432/C58K54, 2006.

[148] E.-L. Tan, J. W. Yeow, S. Peksi, H. Li, Z. Yang, and W.-S. Gan, "Sntl-ntu dcase25 submission: Acoustic scene classification using CNN-GRU model without knowledge distillation", DCASE2025 Challenge, Tech. Rep., 2025.

[149] PyTorch Team, *Torch.permute — pytorch 2.2 documentation*, Accessed: 2025-04-13, PyTorch, 2024. [Online]. Available: https://pytorch.org/docs/stable/generated/torch.permute.html.

[150] PyTorch Team, *Torch.nn.conv1d — pytorch 2.2 documentation*, Accessed: 2025-04-13, PyTorch, 2024. [Online]. Available: https://pytorch.org/docs/stable/generated/torch.nn.Conv1d.html#torch.nn.Conv1d.

[151] PyTorch Team, *Torch.nn.relu — pytorch 2.2 documentation*, Accessed: 2025-04-13, PyTorch, 2024. [Online]. Available: https://pytorch.org/docs/stable/generated/torch.nn.ReLU.html#torch.nn.ReLU.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[152] PyTorch Team, *Torch.nn.maxpool1d — pytorch 2.2 documentation*, Accessed: 2025-04-13, PyTorch, 2024. [Online]. Available: https://pytorch.org/docs/stable/generated/torch.nn.MaxPool1d.html#torch.nn.MaxPool1d.

[153] PyTorch Team, *Torch.nn.lstm — pytorch 2.2 documentation*, Accessed: 2025-04-13, PyTorch, 2024. [Online]. Available: https://pytorch.org/docs/stable/generated/torch.nn.LSTM.html#torch.nn.LSTM.

[154] PyTorch Team, *Torch.nn.linear — pytorch 2.2 documentation*, Accessed: 2025-04-13, PyTorch, 2024. [Online]. Available: https://pytorch.org/docs/stable/generated/torch.nn.Linear.html#torch.nn.Linear.

[155] PyTorch Team, *Torch.nn.flatten — pytorch 2.2 documentation*, Accessed: 2025-04-13, PyTorch, 2024. [Online]. Available: https://pytorch.org/docs/stable/generated/torch.nn.Flatten.html#torch.nn.Flatten.

[156] PyTorch Team, *Torch.nn.lazylinear — pytorch 2.2 documentation*, Accessed: 2025-04-13, PyTorch, 2024. [Online]. Available: https://pytorch.org/docs/stable/generated/torch.nn.LazyLinear.html#torch.nn.LazyLinear.

[157] H. J. Sadaei, P. C. de Lima e Silva, F. G. Guimarães, and M. H. Lee, "Short-term load forecasting by using a combined method of convolutional neural networks and fuzzy time series", *Energy*, vol. 175, pp. 365–377, 2019, ISSN: 0360-5442. DOI: 10.1016/j.energy.2019.03.081. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360544219304852.

[158] Y. Luo, H. Liu, L. Shi, and L. Gan, "Dynacp: Dynamic parallel selective convolution in cp-mobile under multi-teacher distillation for acoustic scene classification", DCASE2025 Challenge, Tech. Rep., 2025.

[159] PyTorch Team, *Torch.cat — pytorch 2.2 documentation*, Accessed: 2025-04-13, PyTorch, 2024. [Online]. Available: https://pytorch.org/docs/stable/generated/torch.cat.html.

[160] PyTorch Team, *Torchvision.models.resnet34 — torchvision main documentation*, Accessed: 2025-04-13, PyTorch, 2024. [Online]. Available: https://docs.pytorch.org/vision/main/models/generated/torchvision.models.resnet34.html.

[161] PyTorch Team, *Resnet34_weights — torchvision main documentation*, Accessed: 2025-04-13, PyTorch, 2024. [Online]. Available: https://docs.pytorch.org/vision/main/_modules/torchvision/models/resnet.html#ResNet34_Weights.

[162] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", *arXiv preprint arXiv:1512.03385*, 2015. arXiv: 1512.03385 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1512.03385.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[163] S. Zagoruyko and N. Komodakis, "Wide residual networks", *arXiv preprint arXiv:1605.07146*, 2017. arXiv: 1605.07146 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1605.07146.

[164] I. Bello, W. Fedus, X. Du, *et al.*, "Revisiting resnets: Improved training and scaling strategies", *arXiv preprint arXiv:2103.07579*, 2021. arXiv: 2103.07579 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2103.07579.

[165] D. Karasin, I.-C. Olariu, M. Schöpf, and A. Szymańska, "Domain-specific external data pre-training and device-aware distillation for data-efficient acoustic scene classification", DCASE2025 Challenge, Tech. Rep., 2025.

[166] PyTorch Team, *Torch.tensor.backward — pytorch 2.2 documentation*, Accessed: 2025-04-13, PyTorch, 2024. [Online]. Available: https://docs.pytorch.org/docs/stable/generated/torch.Tensor.backward.html.

[167] PyTorch Team, *Torch.nn.module — pytorch 2.2 documentation*, Accessed: 2025-04-13, PyTorch, 2024. [Online]. Available: https://docs.pytorch.org/docs/stable/generated/torch.nn.Module.html#torch.nn.Module.

[168] PyTorch Team, *Torch.nn.crossentropyloss — pytorch 2.2 documentation*, Accessed: 2025-04-13, PyTorch, 2024. [Online]. Available: https://docs.pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html.

[169] PyTorch Team, *Torch.optim.adam*, Online documentation, Accessed: 2025-04-13, 2024. [Online]. Available: https://pytorch.org/docs/stable/generated/torch.optim.Adam.html.

[170] PyTorch Team, *Torch.utils.data.dataloader — pytorch 2.2 documentation*, Accessed: 2025-06-23, PyTorch, 2024. [Online]. Available: https://docs.pytorch.org/docs/stable/data.html#torch.utils.data.DataLoader.

[171] scikit-learn developers, *Sklearn model-selection - train-test-split — scikit-learn-1.4.2 documentation*, Accessed: 2025-06-23, scikit-learn, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html.

[172] scikit-learn developers, *Stratification — scikit-learn 1.4.2 documentation*, Accessed: 2025-06-23, scikit-learn, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/cross_validation.html#stratification.

[173] tslearn developers, *Tslearn.preprocessing.timeseriesscalerminmax — tslearn 0.6.2 documentation*, Accessed: 2025-06-23, tslearn, 2024. [Online]. Available: https://tslearn.readthedocs.io/en/latest/gen_modules/preprocessing/tslearn.preprocessing.TimeSeriesScalerMinMax.html.

[174] PyTorch Team, *Torch.utils.data.dataset — pytorch 2.2 documentation*, Accessed: 2025-06-23, PyTorch, 2024. [Online]. Available: https://docs.pytorch.org/docs/stable/data.html#torch.utils.data.Dataset.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[175] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", *arXiv preprint arXiv:1412.6980*, 2017. arXiv: 1412.6980 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1412.6980.

[176] C. E. Tripp, J. Perr-Sauer, J. Gafur, *et al.*, *Measuring the energy consumption and efficiency of deep neural networks: An empirical analysis and design recommendations*, 2024. arXiv: 2403.08151 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2403.08151.

[177] PyTorch Team, *Torch.optim.optimizer.zero*$_g$*rad*$^-$*pytorch*2.2*documentation*, Accessed: 2025-07-03, PyTorch, 2024. [Online]. Available: https://docs.pytorch.org/docs/stable/generated/torch.optim.Optimizer.zero_grad.html.

[178] PyTorch Team, *Torch.optim.optimizer.step — pytorch 2.2 documentation*, Accessed: 2025-07-03, PyTorch, 2024. [Online]. Available: https://docs.pytorch.org/docs/stable/generated/torch.optim.Optimizer.step.html.

[179] PyTorch Team, *Torch.save — pytorch 2.2 documentation*, Accessed: 2025-07-03, PyTorch, 2024. [Online]. Available: https://docs.pytorch.org/docs/stable/generated/torch.save.html.

[180] PyTorch Team, *Torch.max — pytorch 2.2 documentation*, Accessed: 2025-07-03, PyTorch, 2024. [Online]. Available: https://docs.pytorch.org/docs/stable/generated/torch.max.html.

[181] D. Torbunov, Y. Huang, H. Yu, *et al.*, "Uvcgan: Unet vision transformer cycle-consistent gan for unpaired image-to-image translation", 2022. arXiv: 2203.02557 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2203.02557.

[182] R. Dautov, H. Song, D. Roman, *et al.*, "Intend: Human-like intelligence for intent-based data operations in the cognitive computing continuum", eng, 2024. [Online]. Available: https://hdl.handle.net/20.500.12708/211167.

[183] I. A. Kapetanidou, A. Nizamis, E. Karanastasis, *et al.*, "Cognitive computing continuum: State-of-the-art review and enact vision & approach", *Journal of Grid Computing*, vol. 23, no. 3, p. 23, 2025, ISSN: 1572-9184. DOI: 10.1007/s10723-025-09810-9. [Online]. Available: https://doi.org/10.1007/s10723-025-09810-9.

[184] D. N. Karger, Y. Chauvier, and N. E. Zimmermann, "Chelsa-cmip6 1.0: A python package to create high resolution bioclimatic variables based on chelsa ver. 2.1 and cmip6 data", *Ecography*, vol. 2023, no. 6, e06535, 2023. DOI: 10.1111/ecog.06535. [Online]. Available: https://nsojournals.onlinelibrary.wiley.com/doi/abs/10.1111/ecog.06535.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[185] A. Paz, T. Lauber, T. W. Crowther, and J. van den Hoogen, "Improving access and use of climate projections for ecological research through the use of a new python tool", *Ecography*, vol. 2024, no. 5, e07186, 2024. DOI: 10.1111/ecog.07186. [Online]. Available: https://nsojournals.onlinelibrary.wiley.com/doi/abs/10.1111/ecog.07186.

[186] M. Bramer, *Principles of Data Mining* (Undergraduate Topics in Computer Science), 4th ed. 2020. Springer London, 2020, ISBN: 1447174933. [Online]. Available: https://doi.org/10.1007/978-1-4471-7493-6.

[187] M. J. Islam, S. Ahmad, F. Haque, M. B. I. Reaz, M. A. S. Bhuiyan, and M. R. Islam, "Application of min-max normalization on subject invariant emg pattern recognition", *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–1, 2022. [Online]. Available: https://doi.org/10.1109/TIM.2022.3154627.

[188] J. Liang, I. Nolasco, B. Ghani, H. Phan, E. Benetos, and D. Stowell, *Mind the domain gap: A systematic analysis on bioacoustic sound event detection*, 2024. arXiv: 2403.18638 [eess.AS]. [Online]. Available: https://arxiv.org/abs/2403.18638.

[189] J. Liang, I. Nolasco, B. Ghani, H. Phan, E. Benetos, and D. Stowell, "Mind the domain gap: A systematic analysis on bioacoustic sound event detection", eng, in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2024, pp. 1257–1261, ISBN: 9789464593617. DOI: 10.23919/eusipco63174.2024.10714948.

[190] I. Nolasco, S. Singh, V. Morfi, *et al.*, "Learning to detect an animal sound from five examples", *Ecological Informatics*, vol. 77, p. 102 258, 2023, ISSN: 1574-9541. DOI: 10.1016/j.ecoinf.2023.102258. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S157495412300287X.

[191] E. Schall, I. I. Kaya, E. Debusschere, P. Devos, and N. Parceritas, "Deep learning in marine bioacoustics: A benchmark for baleen whale detection", *Remote Sensing in Ecology and Conservation*, vol. 10, no. 5, pp. 642–654, 2024. DOI: 10.1002/rse2.392. eprint: https://zslpublications.onlinelibrary.wiley.com/doi/pdf/10.1002/rse2.392. [Online]. Available: https://zslpublications.onlinelibrary.wiley.com/doi/abs/10.1002/rse2.392.

[192] S. Umesh, L. Cohen, and D. Nelson, "Fitting the mel scale", in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Apr. 1999, pp. 217–220, ISBN: 0-7803-5041-3. DOI: 10.1109/ICASSP.1999.758101.

[193] O. S. Initiative, *The mit no attribution license (mit-0)*, Website, Accessed: 2025-04-13, 2020. [Online]. Available: https://opensource.org/license/mit-0.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[194] J. Brownlee, *Pytorch ultimate 2024*, Online course, Accessed: 2025-04-13, 2024. [Online]. Available: https://learning.oreilly.com/course/pytorch-ultimate-2024/9781801070089/.

[195] O. Fagbohun, *TensorFlow Developer Certificate Guide: Efficiently Tackle Deep Learning and ML Problems to Ace the Developer Certificate Exam*, English, 1st. Birmingham, UK: Packt Publishing, 2023, ISBN: 9781803249204.

[196] R. Fletcher and J. Breitling, "Market mechanism or subsidy in disguise? governing payment for environmental services in costa rica", *Geoforum*, vol. 43, no. 3, pp. 402–411, 2012, The Global Rise and Local Implications of Market-Oriented Conservation Governance, ISSN: 0016-7185. DOI: https://doi.org/10.1016/j.geoforum.2011.11.008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0016718511002107.

[197] M. R. W. Rands, W. M. Adams, L. Bennun, *et al.*, "Biodiversity conservation: Challenges beyond 2010", *Science*, vol. 329, no. 5997, pp. 1298–1303, 2010. DOI: 10.1126/science.1189138. eprint: https://www.science.org/doi/pdf/10.1126/science.1189138. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1189138.

[198] Z. K. Abdul and A. K. Al-Talabani, "Mel frequency cepstral coefficient and its applications: A review", *IEEE Access*, vol. 10, pp. 122 136–122 158, 2022. DOI: 10.1109/ACCESS.2022.3223444.

[199] S. Choudhary, C. R. Karthik, P. S. Lakshmi, and S. Kumar, "Lean: Light and efficient audio classification network", in *2022 IEEE 19th India Council International Conference (INDICON)*, 2022, pp. 1–6. DOI: 10.1109/INDICON56171.2022.10039921.

[200] S. Abu-El-Haija, N. Kothari, J. Lee, *et al.*, "Youtube-8m: A large-scale video classification benchmark", *CoRR*, vol. abs/1609.08675, 2016. arXiv: 1609.08675. [Online]. Available: http://arxiv.org/abs/1609.08675.

[201] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, *et al.*, "Audio set: An ontology and human-labeled dataset for audio events", in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780. DOI: 10.1109/ICASSP.2017.7952261.

[202] O. C. Metcalf, J. Barlow, C. Devenish, S. Marsden, E. Berenguer, and A. C. Lees, "Acoustic indices perform better when applied at ecologically meaningful time and frequency scales", *Methods in Ecology and Evolution*, vol. 12, no. 3, pp. 421–431, 2021. DOI: 10.1111/2041-210X.13521. eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13521. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13521.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[203] O. Sánchez-Chaves and G. Navarrete-Chacón, "La experiencia de costa rica en el pago por servicios ambientales: 20 años de lecciones aprendidas", *Revista de Ciencias Ambientales*, vol. 51, no. 2, pp. 195–214, 2017. DOI: 10.15359/rca.51-2.11. [Online]. Available: https://www.revistas.una.ac.cr/index.php/ambientales/article/view/9487.

[204] A. Kershenbaum, Akçay, L. Babu-Saheer, *et al.*, "Automatic detection for bioacoustic research: A practical guide from and for biologists and computer scientists", *Biological Reviews*, vol. n/a, no. n/a, DOI: 10.1111/brv.13155. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/brv.13155. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/brv.13155.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.