AHMAD QURTHOBI

DEEP NEURAL NETWORK-BASED METHOD FOR DETECTING ANOMALY EVENTS IN NOISY ACOUSTIC ENVIRONMENTS

DOCTORAL DISSERTATION

Kaunas
2026

KAUNAS UNIVERSITY OF TECHNOLOGY

AHMAD QURTHOBI

DEEP NEURAL NETWORK-BASED METHOD FOR DETECTING ANOMALY EVENTS IN NOISY ACOUSTIC ENVIRONMENTS

Doctoral Dissertation
Technological Science, Informatics Engineering (T 007)

2026, Kaunas

The dissertation has been prepared at the Department of Software Engineering of the Faculty of Informatics of Kaunas University of Technology in 2021–2025.

The doctoral right has been granted to Kaunas University of Technology together with Vilnius Gediminas Technical University.

Research supervisor:
Prof. Dr. Rytis MASKELIŪNAS (Kaunas University of Technology, Technological Sciences, Informatics Engineering, T 007).

Edited by: English language editor Dr. Armandas Rumšas (Publishing House Technologija), Lithuanian language editor Aurelija Gražina Rukšaitė (Publishing House Technologija).

Dissertation Defence Board of Informatics Engineering Science Field:
Prof. Dr. Renaldas URNIEŽIUS (Kaunas University of Technology, Technological Sciences, Informatics Engineering, T 007) – **chairperson**;
Prof. Dr. Nikolaj GORANIN (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering, T 007);
Assoc. Prof. Dr. Zenun KASTRATI (Linnaeus University, Sweden, Technological Sciences, Informatics Engineering, T 007);
Prof. Dr. Renaldas RAIŠUTIS (Kaunas University of Technology, Technological Sciences, Measurement Engineering, T 010);
Prof. Dr. Dmitrij ŠEŠOK (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering, T 007).

The dissertation defence will be held on 28 April 2026, at 10:30 a.m. at the Rectorate Hall of Kaunas University of Technology in the meeting of the Dissertation Defence Board of the Informatics Engineering Science Field.

Address: K. Donelaičio 73-402, LT-44249 Kaunas, Lithuania.
Phone: +370 608 28 527; e-mail doktorantura@ktu.lt

The dissertation was sent out on 27 March 2026.
The dissertation is available on the website http://ktu.edu, at the library of Kaunas University of Technology (Gedimino 50, Kaunas) and Vilnius Gediminas Technical University (Saulėtekio 14, Vilnius).

KAUNO TECHNOLOGIJOS UNIVERSITETAS

AHMAD QURTHOBI

GILIAIS NEURONŲ TINKLAIS PAGRĮSTAS GARSO ANOMALIJŲ APTIKIMO TRIUKŠMINGOJE AKUSTINĖJE APLINKOJE METODAS

Daktaro disertacija
Technologikos mokslai, informatikos inžinerija (T 007)

2026, Kaunas

Disertacija rengta 2021–2025 metais Kauno technologijos universiteto Informatikos fakultete, Programų inžinerijos katedroje.

Doktorantūros teisė Kauno technologijos universitetui suteikta kartu su Vilniaus Gedimino technikos universitetu.

Mokslinis vadovas:
prof. dr. Rytis MASKELIŪNAS (Kauno technologijos universitetas, technologijos mokslai, informatikos inžinerija, T 007).

Redagavo: anglų kalbos redaktorius dr. Armandas Rumšas (leidykla „Technologija“), lietuvių kalbos redaktorė Aurelija Gražina Rukšaitė (leidykla „Technologija“).

Informatikos inžinerijos mokslo krypties disertacijos gynimo taryba:
prof. dr. Renaldas URNIEŽIUS (Kauno technologijos universitetas, technologijos mokslai, informatikos inžinerija, T 007) – **pirmininkas**;
prof. dr. Nikolaj GORANIN (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija, T 007);
doc. dr. Zenun KASTRATI (Linėjaus universitetas, Švedija, technologijos mokslai, informatikos inžinerija, T 007);
prof. dr. Renaldas RAIŠUTIS (Kauno technologijos universitetas, technologijos mokslai, matavimų inžinerija, T 010);
prof. dr. Dmitrij ŠEŠOK (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija, T 007).

Disertacija bus ginama viešame Informatikos inžinerijos mokslo krypties disertacijos gynimo tarybos posėdyje 2026 m. balandžio 28 d. 10:30 val. Kauno technologijos universiteto Rektorato salėje.

Adresas: K. Donelaičio g. 73-402, LT-44249 Kaunas, Lietuva.
Tel: +370 608 28 527; el. paštas doktorantura@ktu.lt

Disertacija išsiųsta 2026 m. kovo 27 d.
Su disertacija galima susipažinti interneto svetainėje http://ktu.edu, Kauno technologijos universiteto bibliotekoje (Gedimino g. 50, Kaunas) ir Vilniaus Gedimino technikos universiteto bibliotekoje (Saulėtekio al. 14, Vilnius).

TABLE OF CONTENTS

LIST OF TABLES

LENTELIŲ SĄRAŠAS

LIST OF FIGURES

PAVEIKSLŲ SĄRAŠAS

LIST OF ABBREVIATIONS, TERMS, AND SYMBOLS

Abbreviations:

AAD – acoustic anomaly detection;
ACDNet – adaptively combined dilated convolution for monocular panorama depth estimation;
ACM – audio classification method;
AE – auto-encoder;
AI – artificial intelligence;
API – application programming interface;
AST – audio spectrogram transformers;
AUC – area under the receiver operation characteristic curve;
BiGRU – bidirectional-GRU;
BiLSTM – bidirectional-LSTM;
BiRNN – bidirectional-RNN;
BST – broadcast shifted-windows transformers;
CAT – causal audio transformers;
CBAM – convolutional block attention module;
CL-Transformer – contrastive learning-based audio spectrogram transformer;
CLSTM – convolutional long short-term memory;
CNN – convolutional neural network;
CPU – central processing unit;
CUDA – compute unified device architecture;
CV – cross-validation;
dB – decibel;
DCT – discrete cosine transform;
DDCNN – dual-input dual-channel convolutional neural network;
DL – deep learning;
DNN – deep neural network;
DT – decision tree;
DTL – deep transfer learning;
e.g. – *exempli gratia*;
ECA – efficient channel attention;
EffNet-BiGRU – EfficientNet + BiGRU;
EffNet-BiLSTM – EfficientNet + BiLSTM;
EMD – empirical mode decomposition;
ERT – extremely randomized trees;
ESC – environmental sound classification;
ESC-NAS – environmental sound classification using hardware-aware neural architecture search;
et al. – *et alia*;

etc. – *et cetera*;
FFT – fast Fourier transform;
FT – Fourier transform;
GB – gigabyte;
Gb – gigabit;
GHz – gigahertz;
GMM – Gaussian mixture model;
GPU – graphics processing unit;
GRU – gated-recurrent unit;
GS – gammatone-spectrogram;
HMM – hidden Markov model;
Hz – hertz;
i.e. – *id est*;
IEMD – incomplete empirical mode decomposition;
kNN – *k*-nearest neighborhood;
LEAN – light and efficient audio classification network;
LSTM – long short-term memory;
MBConv – mobile inverted bottleneck convolution;
MFR – multi-frequency resolution;
ML – machine learning;
MLR – methodological literature review;
NLP – natural language processing;
NLR – narrative literature review;
NN – neural network;
OC-SVM – one-class support vector machine;
OS – operating system;
PANN – pre-trained audio neural networks;
ParallelNet – parallel network;
R&D – research and development;
RAM – random access memory;
ReLU – rectified linear unit;
RF – random forest;
RNN – recurrent neural network;
ROC – receiver operation characteristic;
S3T – self-supervised pre-training with shifted-windows transformers;
SE – squeeze and excitation;
SER – speech emotion recognition;
SLR – systematic literature review;
SNR – signal-to-noise ratio;
SOTA – state of the art;
SPA – spectral pooling attention;
SSD – solid-state drive;

SSKD – self-supervision with knowledge distillation;
SVM – support vector machine;
SWin – shifted-window;
SwinEmoNet – shifted window transformer emotion network;
SWinT-BiGRU – shifted-windows transformers + BiGRU;
SWinT-BiLSTM – shifted-windows transformers + BiLSTM;
TB – terabyte;
TFECN – time-frequency enhanced ConvNet;
TinyML – tiny machine learning;
TL – transfer learning;
TLR – traditional literature review;
UAV – unmanned aerial vehicle;

Terms:

A4 – musical note A in the fourth octave.
AudioSet – a comprehensive collection of audio event data sourced from YouTube videos and compiled by Google.
BirdCLEF – An internationally recognized annual Kaggle competition on machine learning for automatic acoustic identification of bird species from audio recordings, named as a portmanteau of "Bird Classification, Localization, and Evaluation Forum".
chroma-STFT – a common lossy audio feature that compresses the full frequency spectrum of a signal into 12 bins, each corresponding to one of the twelve pitch classes of the Western scale. The acronym stands for "chroma short term Fourier transform".
ConvNeXt – a set of pure CNN models, released by Meta AI Research in 2022, that update classic convolutional networks by adopting key design ideas from ViT.
DCASE – a publicly available, systematically curated audio dataset designed to benchmark and compare machine learning models for audio analysis tasks, whose name stands for "Detection and Classification of Acoustic Scenes and Events".
DenseNet – stands for "Densely Connected Convolutional Network", a CNN design where each layer links to all later layers within a dense block, enhancing information flow and feature reuse.
EffNet – a group of CNN models and scaling techniques built to deliver top-tier accuracy with minimal computation, collectively referred to as "EfficientNet".
EFSC-24 – a forest sound classification dataset, referred to as the "Enhanced Forest Sound Classification Dataset (EFSC-24)" in a 2021 study by Ahmad et al..
Emo-DB – a widely used public German emotional speech corpus from the Institute of Communication Science at the Technical University of Berlin, commonly referred to as the "Berlin Database of Emotional Speech".
ESC-10 – a subset of the ESC-50 dataset that is restricted to only ten selected classes.
ESC-50 – a dataset of 2,000 environmental audio clips across 50 categories, created as a standard benchmark for environmental sound classification methods.
FSC22 – a public benchmark dataset for detecting and classifying forest audio,

particularly sounds linked to illegal activities like poaching and logging. "FSC" stands for "Forest Sound Classification," and "22" indicates the year of creation.

FSD50K – an open collection of human-annotated sound events comprising 51,197 Freesound audio clips, adding up to 108.3 hours of audio, organized into 200 categories derived from the AudioSet Ontology.

GAN – an abbreviation for "Generative Adversarial Network," a machine learning framework introduced by Ian Goodfellow et al. in 2014 to generate synthetic data similar to a training set.

GoogLeNet – a 22-layer CNN model that achieved first place in the 2014 ImageNet Large-Scale Visual Recognition Challenge.

GTZAN – a widely used public dataset in Music Information Retrieval (MIR) for automatic music genre recognition.

***k*-folds CV** – a widely used resampling technique in machine learning that evaluates a model's performance and generalization capability, short for "*k*-fold cross-validation".

mel-spectrogram – a time–frequency representation of audio that replaces the linear Hertz scale with the non-linear mel scale, reflecting human pitch perception.

MFCC – a set of characteristics that represent the short-term power spectrum of a sound, based on the ear's non-linear, logarithmic way of perceiving frequency. MFCC stands for "Mel-Frequency Cepstral Coefficients".

MIMII – a publicly available collection of industrial machine audio created for unsupervised detection of abnormal sounds. its name is an acronym for "Malfunctioning Industrial Machine Investigation and Inspection".

MIMII-DUE – a sound dataset for malfunctioning industrial machine investigation and inspection with domain shifts due to changes in operational and environmental conditions.

MobileNet – a set of compact, high-efficiency CNN models by Google, optimized for vision tasks on mobile and edge devices.

RAVDESS – an acronym for "Ryerson Audio-Visual Database of Emotional Speech and Song," a validated multimodal dataset widely used in ML, psychology, and SER.

ResNet – a short for "Residual Network", a landmark deep CNN model proposed by He et. al. at Microsoft Research in 2015.

STFT – an abbreviation of "Short-Time Fourier Transform," a mathematical method for analyzing how the frequency and phase content of short segments of a signal vary over time.

SWinT – a shortened form of "Shifted Window Transformer", a hierarchical vision transformer backbone designed for broad computer vision tasks, including classification, detection, and segmentation.

t-SNE – an abbreviation of "t-distributed Stochastic Neighbor Embedding," a nonlinear dimensionality reduction technique mainly used to project high-dimensional data into 2D or 3D for visualization.

ToyADMOS – a collection of operating sound recordings from small-scale machines for abnormal sound detection.

UrbanSound8k – a publicly accessible audio corpus created for evaluating and comparing machine learning models on the task of environmental sound classification.
VGG – an abbreviation of "Visual Geometry Group," a seminal deep CNN model proposed by Karen Simonyan and Andrew Zisserman (University of Oxford) in their 2014 paper "Very Deep Convolutional Networks for Large-Scale Image Recognition".
ViT – short for "Vision Transformer," a computer vision model that uses a standard Transformer encoder directly on images.
Wi-Fi – short for "Wireless Fidelity," a wireless standard that connects devices to a WLAN and the internet via radio waves.
XGBoost – a shortened form of "eXtreme Gradient Boosting," an open-source library offering a fast, regularized implementation of Gradient Boosted Decision Trees (GBDT).
YAMNet – a Google pre-trained deep learning model based on MobileNetV1 that identifies and categorizes 521 types of audio events (such as speech, music, animal sounds, and ambient noise) using the large-scale AudioSet dataset. Its name is an acronym for "Yet Another Mobile Network".

Symbols:

$\mathtt{D}$ – original dataset;
$\mathtt{D}_k$ – a dataset's fold;
ε – small constant;
f – frequency bin;
f_{mel} – Mel-scaled frequency;
$\mathscr{F}_{-}$ – false negative;
$\mathscr{F}_{+}$ – false positive;
$\mathscr{F}_{+R}$ – false positive rate;
f_{ref} – fixed reference frequency;
F – total numbers of frequency bins;
$H_m[f]$ – Mel-filters weight of frequency;
H – hop length;
$\mathtt{hop\_length}$ – hop length (in Librosa library);
k – an arbitrary integer number;
κ – coefficient of discrete cosine transform;
K – total folds;
m – Mel-filter index;
$\text{MFCC}[t,\kappa]$ – MFCC resulting coefficient;
$\mathscr{M}_k$ – performance metric on the k-th validation fold;
mod 12 – modulo-12;
M – numbers of Mel-fitlers;
$\bar{\mathscr{M}}$ – average performance metric;
n – discrete-time frame index;
$\mathtt{n\_chroma}$ – number of chroma bins to produce;

`n_fft` – FFT windows size;
N – fast Fourier transform size;
`n_mels` – number of mel bands;
`n_mfcc` – number of MFCCs to return;
$\omega[n - tH]$ – sliding-window function;
p – chroma bin;
P_{ref} – reference power value;
$S_{\text{chroma}}[t, p]$ – chroma energy;
$\sigma_{\mathscr{M}}$ – deviation standard of performance metric;
$S_{\mathbf{mag}}[t, f]$ – magnitude spectrogram;
$S_{\text{mel}}[t, m]$ – Mel-spectral function;
$S_{\text{mel,log}}[t, m]$ – log-mel spectrogram;
$S_{\text{mel,norm}}[t, m]$ – mel-power normalization function;
$S_{\text{P}}[t, f]$ – power spectrogram;
`sr` – sampling rate;
t – time;
$\mathscr{T}_{-}$ – true negative;
$\mathscr{T}_{+}$ – true positive;
$\mathscr{T}_{+R}$ – true positive rate;
$x[n]$ – discrete-time input signal;
$X[t, f]$ – spectrogram;

INTRODUCTION AND RESEARCH MOTIVATION

Relevance of the Work

Audio classification plays a crucial role in contemporary acoustic anomaly detection (AAD) systems by providing essential insights suitable for various applications [1–3]. These fields are valid not only for industrial environments [4] but also for dynamic urban areas [5, 6] and even complex natural ecosystems, such as forests [7]. Moreover, collecting audio data from the surroundings can occur without interrupting ongoing operations or cycles [8, 9]. However, each environment offers its unique challenges and opportunities for detecting unusual audio patterns, particularly in contexts where visual information is unavailable or limited. In addition, the interference that accompanies the sound during data collection is an integral component of data collection, although it is often considered a nuisance.

Acquiring clean, noise-free audio in real-world conditions is challenging because of pervasive background sounds. Many studies have demonstrated that such noise substantially degrades sound clarity and classification accuracy [10–12]. Taneja et al. (2013) [13] further point out that creating a truly noise-free environment typically demands complex, expensive setups that are rarely feasible for on-site recordings. In addition, audio interference from humans, animals, machinery, and environmental sources can almost never be fully removed [14, 15]. Such disturbances introduce variability that hinders a model's ability to separate normal from abnormal events, jeopardizing data integrity and lowering the accuracy of downstream analyses. Consequently, AAD systems should incorporate signal conditioning and feature extraction methods to address these challenges.

Under ideal, noise-free conditions, people can naturally perceive their surroundings without specialized tools. In practice, however, such conditions are rare; real environments typically contain disturbances and noise that impair both observation and decision-making [16–18]. Barchiesi et al. (2015) [19] showed that these factors reduce the efficiency, accuracy, and consistency of human perception while complicating classification tasks. Recent advances now allow accurate extraction of audio signals from noisy backgrounds, improving the clarity and reliability of collected data. These methods can be readily embedded in hardware and software, expanding their practical use [20–23]. They rely on the assumption that every source has a unique acoustic signature that reflects its state or behavior. For instance, a fan emits sound in a specific frequency band, and deviations can signal faults or disturbances. This principle has driven improved acoustic filtering and signal analysis techniques, which are vital for audio-based anomaly detection systems used to identify equipment failures, malfunctions, or unusual events in noisy settings. Consequently, audio classification technology is increasingly important for real-time monitoring and automated decision-making, supported by advances in computation and ongoing research.

An expanding body of research shows that advances in computational

technologies, especially in artificial intelligence (AI), machine learning (ML), and deep learning (DL), have greatly enhanced researchers' capacity to make accurate observations with minimal additional refinement [24]. In particular, AI methods based on ML and DL offer powerful and efficient solutions for optimizing classification tasks in complex, noisy settings. DL models can extract and learn sophisticated patterns from large datasets, delivering precise predictions where traditional observational approaches often fail. Nonetheless, their performance critically depends on the quality and heterogeneity of the training data [25]. Consequently, the careful design and curation of comprehensive, well-structured datasets are essential to developing effective AI systems for real-world use.

In audio-based classification, labels or classes are typically defined by sound sources and their characteristic frequency ranges reflecting their operating behavior [26]. This strategy is especially effective for industrial machinery, where each type of machine exhibits distinctive operating signatures (such as frequency spectra, vibration patterns, and even temperature) shaping its acoustic footprint. Under normal operating conditions, these features form a baseline pattern; small deviations in the emitted sound can signal abnormal states and possible performance degradation [27].

Compared to other domains, natural soundscapes require a different analytical approach than structured or human-dominated acoustic settings. Forests and jungles are intricate ecosystems with diverse plant and animal communities, where acoustic activity mirrors biological interactions, environmental processes, and human impact. Examining forest acoustics through faunal diversity yields key insights into ecosystem dynamics and eco-acoustics [28]. Monitoring these habitats involves classifying large numbers of nonstationary overlapping sounds [29–33]. As the number of target sound classes grows, the classification task becomes increasingly complex, demanding careful adaptation of pre-trained models and advanced DL methods. Such models must be tuned to specific recognition goals, e.g., environmental [30], ambient [34, 35], or animal sound classification [36]. To address these challenges, many studies over the last decade have gathered and processed audio captured under realistic, noise-contaminated recording conditions.

A key goal has been to build datasets that both enable ongoing research and accurately reflect real-world acoustic environments. Within this context, Piczak [5] proposed the ESC-50 dataset in 2015, now a standard benchmark of environmental sounds from residential and urban scenes, containing 50 everyday classes (e.g., coughing, breathing, footsteps) sourced from FreeSound.org [37]. Later, Purohit et al. at Hitachi Ltd. released the MIMII dataset [4], targeting industrial machine audio recorded under multiple background noise conditions (i.e., -6, 0, and +6 decibels (dBs)) to emulate normal operation. More recently, Bandara et al. [7] introduced the FSC22 dataset, a large-scale collection of natural soundscapes covering biotic, anthropogenic, and geophysical events for audio-based forest monitoring. Additional resources (e.g., UrbanSound8k [6], AudioSet [17], DCASE [38], ToyADMOS [39],

FSD50K [40], etc.) further expand the landscape from niche environmental audio corpora to broadly applicable benchmarks.

This dissertation studies robust audio-based anomaly detection by assessing deep neural network (DNN) models (focusing on convolutional neural network (CNN) and transformer architectures) across varied acoustic conditions. Industrial, urban, and forest soundscapes are treated as complementary domains, covering structured machinery noise, diverse human activities, and highly variable natural sounds. By using public datasets such as MIMII [4], ESC-50 [5], and FSC22 [7], the work analyzes model generalization under different noise levels, class imbalance, and spectral profiles. Perceptually inspired representations (mel-spectrogram [41, 42], MFCC [42], and chroma-STFT [43–45]) feed hybrid DL models that combine pretrained convolutional or transformer backbones with bidirectional-LSTM (BiLSTM) or bidirectional-GRU (BiGRU) layers to jointly model spatial and temporal structures. A unified k-folds CV setup provides statistically robust and comparable evaluations across all datasets. Here, anomalies are defined as sound events or operating states whose spectro-temporal signatures diverge from expected patterns and are treated as separate classes in a supervised multi-class setting.

Formulation of Problems

Audio-based classification has strong potential in many domains, but recurring issues still obstruct its effective use in real-world data collection and processing. These issues are particularly critical when building robust, generalizable models for anomaly detection in acoustically diverse settings. Instead of treating them only as obstacles, they can be seen as drivers of innovation in data representation, model design, and evaluation methods. The key challenges that continue to shape and constrain progress in this area are:

1. Although handcrafted audio features like MFCC and chroma-based representations are widely used, there is still no systematic understanding of how various feature types interact with modern DL architectures under diverse acoustic conditions. In particular, it remains uncertain which features deliver the most consistent and robust performance when applied to individual datasets that exhibit noise, heterogeneity, and context-dependent audio events.
2. Industrial audio datasets, such as those employed for machine anomaly detection, typically exhibit strong class imbalance and diverse noise conditions, which can substantially reduce model reliability on the given dataset. Many current methods fail to deliver consistent performance in these scenarios, underscoring the demand for robust techniques that address imbalance and noise sensitivity without depending on cross-dataset training.
3. While temporal dependencies are crucial in real-world audio signals, the advantages of adding recurrent layers to modern DL architectures for within-dataset classification are still not well established. In particular, it is unclear when temporal modeling actually improves classification accuracy, and how its impact varies with different feature representations and backbone architectures.

4. Different DL backbone networks, spanning convolutional and transformer-based designs, offer distinct abilities to capture spectral and temporal characteristics. Yet, there is a lack of systematic evidence on how to best align feature selection with architectural choices to optimize classification accuracy when training and testing on the same dataset.
5. Many prior works rely on dataset-specific or otherwise constrained evaluation setups, hindering fair comparison between different feature–model pairings and reducing reproducibility. Standardized evaluation procedures are needed to support robust within-dataset performance assessment across a wide range of acoustic benchmarks.
6. Although leading AAD systems often report strong accuracy, these improvements are frequently accompanied by higher computational demands and less stable training. Achieving an effective trade-off between performance, robustness, and computational efficiency is still unresolved for within-dataset audio classification, especially in practical or resource-limited settings.

Aim of the Research

This research aims to improve the accuracy, robustness, and reliability of AAD in noisy environments. It focuses on better distinguishing normal from abnormal acoustic patterns across varied noise conditions by examining how audio feature representations interact with DNN architectures. To this end, it develops and evaluates supervised DL models that use perceptually relevant features and temporal modeling to reduce the impact of background noise, class imbalance, and acoustic variability within datasets.

Research Objectives

The overarching goal of this study is to enhance the accuracy, robustness, and dependability of detecting anomalous events in noisy acoustic settings through DNN-based approaches. In pursuit of this goal, the research is structured around the following specific objectives:

1. To review and integrate prior work on AAD and environmental sound classification (ESC), focusing on feature representations, DNN models, and evaluation methods in noisy environments.
2. To examine how data imbalance and varying noise conditions affect anomaly-aware audio classification in industrial acoustic settings, and to assess augmentation methods that enhance detection robustness.
3. To evaluate how well various audio feature representations can separate normal from anomalous acoustic patterns in a range of noisy environments.
4. To systematically assess how different DNN architectures (such as CNN-based, transformer-based models, and hybrid temporal approaches) perform in detecting anomalous events, by using supervised learning methods applied to AAD tasks.
5. To systematically evaluate how temporal modeling strategies influence AAD effectiveness, with a specific focus on their interaction with feature representations and the choice of backbone architectures.

6. To systematically examine the balance between anomaly detection accuracy and computational efficiency, while also assessing model robustness and its appropriateness for real-world applications, including those with limited computational or memory resources.

Scientific Novelty

This research introduces several significant scientific contributions that extend the current state of knowledge in AAD and ESC. The key aspects of novelty are outlined below:

1. An important advancement is the full use of the MIMII dataset, covering all devices and signal-to-noise ratio (SNR) levels, unlike past studies that focused on single devices or limited noise conditions [23, 46]. This extensive approach enhances classification by exposing models to varied acoustic patterns. However, effective feature selection and model choice are crucial to balance computational cost and maintain practical feasibility.
2. Mel-spectrogram has outperformed nearly all other traditional descriptors tested, such as MFCC and chroma-STFT in the extraction of characteristics. Its advantage lies in maintaining detailed time–frequency representations that align well with human hearing. However, achieving this comes with significantly higher computational costs: experiments show that training and validation mel-spectrogram takes roughly two to three times longer than other features, highlighting a trade-off between precision and efficiency crucial for real-time or embedded designs. In addition, Figure 19 illustrates that utilization of MFCC outperforms mel-spectrogram's achievement, on average, for the combination with EfficientNet + BiLSTM (EffNet-BiLSTM).
3. The study highlights that combining mel-spectrogram features with shifted-windows transformers + BiLSTM (SWinT-BiLSTM) architectures effectively classifies datasets with limited data, like ESC-50 and FSC22. This pairing outperforms other configurations, making it ideal for heterogeneous, scarce data environments, such as residential or forest areas. While it did not exceed the mel-spectrogram+SWinT baseline for the larger, imbalanced MIMII dataset, the performance gap was slight, which indicates its competitiveness in more challenging industrial scenarios.
4. Integrating recurrent modules into pre-trained architectures is another key contribution. Utilizing recurrent neural network (RNN)-based temporal refinement, especially with BiLSTM layers, delays overfitting more than traditional feed-forward models. This approach improves accuracy and temporal sensitivity but increases model size and training duration, offering insights for optimizing deep temporal models in resource-limited settings.

Practical Significance

1. The application of DL to perform acoustic-based anomaly investigation and identification offers a promising avenue for the development of environmental monitoring methods. The non-invasive and non-destructive nature of audio data collection can significantly assist the surveillance process without encountering or

assembling new problems [47–49]. In fact, this methodology can be applied to a wide range of environmental settings.

2. The use of DL in conjunction with acoustic methods has demonstrated efficacy in applications ranging from wildlife monitoring to industrial environments, facilitating the detection and classification of objects or phenomena. In industrial environments, for example, the attributes of the sound produced tend to be predictable [50, 51]. However, the challenge of exceptionally deafening sounds must be encountered when humans are involved as observers [52]. In contrast, sounds from wild environments tend to be more varied, yet the diversity they produce poses a considerable challenge for computation-based recognition [53–55]. Fortunately, these challenges were ultimately overcome. Indeed, environmental noise, which is often regarded as a nuisance, can be adopted as advantageous. Hence, the implementation of suitable signal processing methodologies further substantiates this assertion.
3. As a result, the integration of DL models and audio extractors tailored to specific requirements has yielded optimal results. Although the outstanding result is of the utmost importance, it is imperative to consider the availability of resources to ensure the optimal functioning of these systems.

Dissertation Statements

1. The major class imbalance issue present in the MIMII dataset has been unraveled by utilizing SNR-based enhancement techniques. This approach has significantly improved the robustness, generalization capabilities, and precision of the model, achieving consistently high average accuracies of $99.06\% \pm 0.02\%$ and $95.27\% \pm 0.20\%$ in F1, particularly in scenarios characterized by high levels of noise for transformer-based models on the MIMII dataset.
2. This dissertation demonstrates that choosing the right audio feature representation is a key design factor for DNN-based anomaly detection in noisy acoustic settings. The experiments show that time–frequency features capturing perceptual and spectral structure, especially mel-spectrograms, yield higher robustness, generalization, and noise stability across datasets and architectures. The MFCC emerges as a slightly lower-performing but efficient and reliable option, whereas tonal features like chroma-STFTs are inherently ill-suited to non-tonal, noise-dominated anomaly detection. Overall, effective detection in complex acoustic scenes depends on feature representations that reflect the statistical and perceptual properties of environmental sounds.
3. This dissertation shows that explicit temporal modeling is crucial for reliable DNN–based anomaly detection in noisy, non-stationary acoustic environments. By incorporating temporal modules such as BiGRU and BiLSTM into pre-trained CNN and transformer backbones, the proposed hybrid models more effectively learn the sequential structure of real-world acoustic signals. The gains in detection accuracy, training stability, and robustness on datasets like MIMII and FSC22 demonstrate that temporal dynamics must be explicitly modeled to obtain

dependable anomaly detection under complex, time-varying noise.

4. The suggested classification framework exhibited notable versatility across various datasets that include industrial, urban, and natural environments. It maintained impressive F1 scores and recall rates, particularly when utilizing SWinT-BiLSTM models paired with spectrogram-based features. The enhanced performance is further corroborated by the distinct clustering observed in the t-SNE visualizations.
5. Examining epochs indicated that implementing early stopping when the validation loss leveled off after 10 epochs effectively avoided overfitting. Models enhanced with recurrent structures, such as BiLSTM, exhibited extended training periods along with improved consistency in performance across various features and datasets.

Scientific Approvals

To validate and support the methodology developed in this doctoral research, five scientific contributions have been produced over the past four years. These consist of two articles published in journals indexed in both Web of Science and Scopus, and three papers included in the proceedings of international conferences. In addition, a comprehensive overview of these publications, including detailed bibliographic information and their specific relationship to the work carried out during the doctoral studies, is provided in Appendix 3.3.8: LIST OF SCIENTIFIC PAPERS AND SCIENTIFIC CONFERENCES.

Articles indexed in Web of Science and Scopus

1. Ahmad Qurthobi, Rytis Maskeliūnas, and Robertas Damaševičius. Detection of mechanical failures in industrial machines using overlapping acoustic anomalies: A systematic literature review. Sensors, 22(10), 2022. ISSN 1424-8220. doi: 10.3390/s22103888.
2. Ahmad Qurthobi, Robertas Damaševičius, Vytautas Barzdaitis, and Rytis Maskeliūnas. Robust forest sound classification using pareto-mordukhovich optimized mfcc in environmental monitoring. IEEE Access, PP:1–1, 01 2025. doi: 10.1109/ACCESS.2025.3535796.

International conference proceedings

1. Ahmad Qurthobi and Rytis Maskeliūnas. The effect of augmentation and filtration on noisy environment's acoustic signals to detect abnormalities in industrial machines based on artificial neural networks. Procedia Computer Science, 220:535–544, 2023. ISSN 1877-0509. doi: 10.1016/j.procs.2023.03.068. The 14th International Conference on Ambient Systems, Networks and Technologies Networks (ANT) and The 6th International Conference on Emerging Data and Industry 4.0 (EDI40).
2. Ahmad Qurthobi and Rytis Maskeliunas. Deep learning and acoustic approach for mechanical failure detection in industrial machinery. Journal of Physics: Conference Series, 2673:012032, 12 2023. doi: 10.1088/1742-6596/2673/1/012032.

3. Robertas Damasevicius, Ahmad Qurthobi, and Rytis Maskeliunas. A hybrid machine learning model for forest wildfire detection using sounds. In 2024 19th Conference on Computer Science and Intelligence Systems (FedCSIS), pages 99–106, 2024. doi: 10.15439/2024F7263.

Dissertation Structure

This dissertation is structured into five chapters. Chapter INTRODUCTION AND RESEARCH MOTIVATION presents the research problem and its motivation, defines the concept of acoustic anomalies adopted in this work, and provides a concise rationale for choosing industrial, urban, and forest soundscapes as complementary evaluation domains. Chapter 1 offers an in-depth literature review, covering both foundational and recent developments in the application of AI, ML, and DL techniques to audio-based recognition and classification. These advances underpin the detection of events, identification of anomalies, and recognition of potential threats in a range of environments, including industrial, residential, and forest contexts. Chapter 2 details the datasets and experimental framework used in this research. It introduces the selected audio corpora (MIMII, ESC-50, and FSC22), examines their acoustic properties and relevance, and describes the time–frequency feature representations adopted, namely mel-spectrogram, MFCC, and chroma-STFT. The chapter also presents the proposed hybrid models, which are derived from CNN- and attention-based DNN models, along with the validation protocol, experimental pipeline, and evaluation procedures followed throughout the study. Chapter 3 reports a detailed experimental assessment of the proposed and baseline DL models based on training and validation outcomes. Model performance is quantified using several metrics, including accuracy, area under the receiver operation characteristic curve (AUC), precision, recall, and F1-score. Finally, Chapter CONCLUSIONS AND FUTURE WORKS closes the dissertation by interpreting and critically discussing the main findings and by highlighting avenues and prospects for future research.

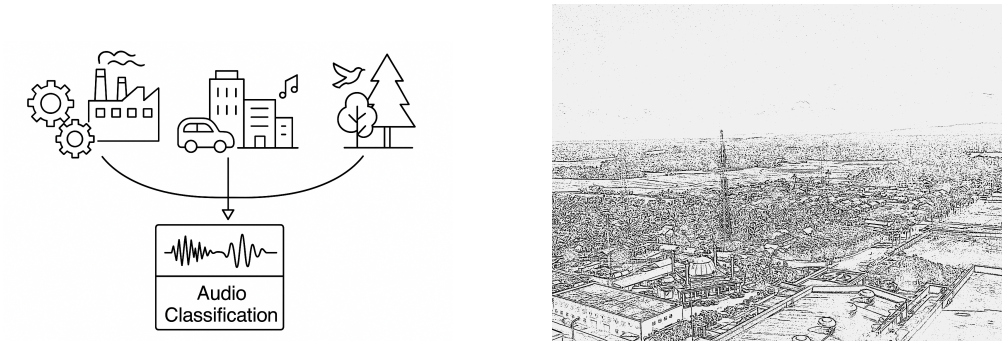
1. LITERATURE REVIEWS OF DEEP-NEURAL NETWORK METHOD FOR ACOUSTIC ANOMALY EVENTS DETECTION

This chapter presents the evolution of research on DNN in acoustics and its application to event detection, environmental monitoring, etc. Section 1.1 examines the progression of audio usage in AI, focusing on time-frequency feature extraction methods and DNN learning models based on CNN and attention mechanisms. Section 1.2 outlines current applications, analyzes the impact of noise, and defines the notion of abnormal conditions adopted in this work. Section 1.3 reviews the literature on acoustic-based anomaly detection, while Section 1.4 discusses recent studies in this domain, particularly in industrial, residential, and natural environments. Finally, Section 1.5 closes the chapter with a summary and the motivation underlying this study.

1.1. Audio Feature Extraction and Deep Neural Networks for Acoustic Classification

A combined analysis of advances in AI methods and new datasets is essential for reliable assessment. Fundamental neural network (NN) architectures (such as long short-term memory (LSTM), gated-recurrent unit (GRU), auto-encoder (AE), and CNN) underpin most modern DL-based classifiers. Since its introduction, LSTM has been extensively applied to classification, forecasting, and control, where it excels at modeling complex temporal dependencies in the input. This capability is particularly important in finance, healthcare, and natural language processing (NLP) [56–62], and it also enables robust audio signal classification by leveraging characteristic signal properties. By comparison, GRU offers a more compact DL layer with fewer parameters and a dedicated forget gate, as described by Cho et al. [63]. Both LSTM and GRU originate from the RNN family. Augmenting them with a bidirectional-RNN (BiRNN) layer yields BiGRU and BiLSTM, which further extend their representational power.

The AE has recently become a prominent approach for acoustic signal classification [64–67]. As a DNN model, it can simultaneously handle classification, denoising, and reconstruction, reproducing the input at the output with high accuracy. Its structure usually consists of simple linear layers arranged into an encoder for compression and a decoder for decompression. In parallel, deep CNN-based NN architectures originally developed for computer vision can be adapted for audio classification [68, 69]. Many widely used pre-trained DL models (e.g., DenseNet [70], EffNet [71, 72], MobileNet [73, 74], and ConvNeXt [75]) are primarily built on CNN. More recently, attention-based architectures have also gained prominence in audio tasks, motivated by the success of CNN-centric methods. Consequently, attention-driven models like SWinT [76, 77] and ViT [78] are being investigated with the objective to enhance classification accuracy further. Yet, all these approaches presuppose a precise and stable signal extraction stage. Despite the advances in deep neural architectures, their performance still critically hinges on the quality and appropriateness of the input feature representations.



(a) sketch **(b)** real environment

Fig. 1. Illustration of the implementation of audio-driven classification

These developments unlock the potential to broaden the scope of audio classification, enabling both the detection of specific anomalies in varied applications and the identification of unexpected events in real environments. Figure 1 provides an example of the extensive applicability of current audio classification approaches. In 2022, Zada et al. [79] argued that environmental noise in existing datasets can function as a useful auxiliary feature rather than a mere artifact. Building on this, Jaiswal and Provost (2023) [80] showed that deliberately injecting noise can improve recognition accuracy. Instead of filtering out such signals, they can be exploited to boost performance, particularly for mitigating class imbalance in binary and multiclass audio classification.

1.1.1. Time–Frequency Audio Representations

In the realm of audio-based classification, the selection of the appropriate audio feature extractor holds as much significance as the choice of classifier model. The general consensus in the field points to STFT, mel-spectrogram, and MFCC as optimal options. Sejdic et al. (2009) [81] review the literature and highlight STFT as the preferred approach to depict sinusoidal signals in the time-frequency domain. This method involves segmenting long, continuous signals into shorter parts, to which the Fourier transform (FT) is subsequently applied. The objective is to analyze the distribution of the Fourier spectrum within each segment, especially considering that the original signal is extensive and randomly arranged. However, employing a straightforward STFT for audio classification can be resource-intensive in terms of computational demands. Consequently, mel-spectrogram and MFCC have emerged as viable alternatives to mitigate these challenges.

The mel-spectrogram and MFCC are widely used techniques that transform audio signals from the time domain into a time–frequency representation. Both methods diverge from the conventional STFT-based approach by employing the Mel scale on the frequency axis rather than a linear frequency scale. This design choice

is motivated by the Mel scale's alignment with human auditory perception, which exhibits non-linear sensitivity to frequency variations. The concept of the Mel scale was first introduced by Stevens et al. in 1937 [41], who demonstrated that perceived pitch more closely follows a perceptually motivated scale than a linear frequency axis. As a result, both mel-spectrogram and MFCC represent audio signals as short-time spectral power distributions warped according to the Mel frequency scale, thereby emphasizing perceptually relevant frequency bands [82]. Building on this foundation, Davis and Mermelstein [42] introduced the Mel filterbank into speech analysis, leading to the development of MFCC, where the application of the discrete cosine transform (DCT) serves to decorrelate spectral coefficients and produce a compact representation suited for statistical modeling. In contrast, chroma-STFT adopts a different perceptual abstraction by mapping spectral energy onto chroma bins that correspond to pitch classes, effectively folding frequencies across octaves [83]. This octave-invariant representation is particularly advantageous for tonal and harmonic audio, such as music, where pitch relationships are more salient than absolute frequency content. Due to its reduced dimensionality and reliance on harmonic structure, chroma-STFT is often associated with lower computational and hardware requirements, although it may discard broadband spectral information that is relevant for non-tonal or noise-dominated sound events.

While MFCC and chroma-STFT were originally designed to address specific perceptual and computational constraints, mel-spectrograms have emerged as a more suitable representation for modern DNNs. Historically, MFCC was developed for speech recognition systems based on generative or shallow discriminative models, where compactness, coefficient decorrelation, and low computational complexity were critical [42]. Similarly, chroma-STFT was introduced to support music information retrieval tasks by emphasizing harmonic relationships and octave invariance, i.e., properties that are beneficial for tonal analysis but less expressive for broadband or transient-dominated signals [83, 84]. In contrast, mel-spectrograms retain dense spectral–temporal structure without enforcing decorrelation or harmonic folding, while preserving local continuity across both time and frequency. This dense representation aligns well with the inductive biases of convolutional and attention-based neural networks, which are designed to exploit spatial locality, hierarchical patterns, and contextual dependencies [5, 85, 86]. Furthermore, the higher dimensionality of mel-spectrograms enables deep models to learn task-specific abstractions directly from data, rather than relying on handcrafted compression stages. As a result, mel-spectrograms have consistently demonstrated superior performance in a wide range of audio classification and anomaly detection tasks, particularly in noisy and heterogeneous environments, where fine-grained spectral cues and temporal dynamics play a crucial role [4, 87, 88].

1.1.2. Convolutional- and Transformer-based Architectures

In recent times, NN has emerged as a leading method in audio classification due to its ability to self-learn complex layered representations from either unprocessed or pre-processed audio data. Architectures based on CNN and attention mechanisms have shown remarkable success when used with time–frequency representations like mel-spectrograms, log-magnitude spectrograms, and chromagrams. Within CNN-centric models, EffNet stands out for leveraging compound scaling, which allows for a balanced increase in the network depth, width, and resolution. Models that employ attention, such as SWinT, enhance the capabilities of self-attention by using non-overlapping shifted windows, thus supporting hierarchical feature modeling while maintaining efficiency and locality.

EffNet and SWinT have been effectively tailored for audio analysis tasks through the use of spectrogram inputs. Porwal (2024) [89] illustrated that ensembles of EffNet were highly effective in identifying bird species during the BirdCLEF challenge. Meanwhile, Wang et al. (2023) [90] introduced time-frequency enhanced ConvNet (TFECN), a convolutional model augmented with time–frequency features, which performed comparably to transformer-driven models on standard benchmarks. In the sphere of transformer architectures, Liu et al. (2023) [91] presented the causal audio transformers (CAT), achieving leading results across various datasets, while Zhao et al. (2022) [92] emphasized the success of SWinT in pre-training for music genre classification. Together, these studies highlight the strength of both CNN and transformer-based frameworks in audio classification tasks.

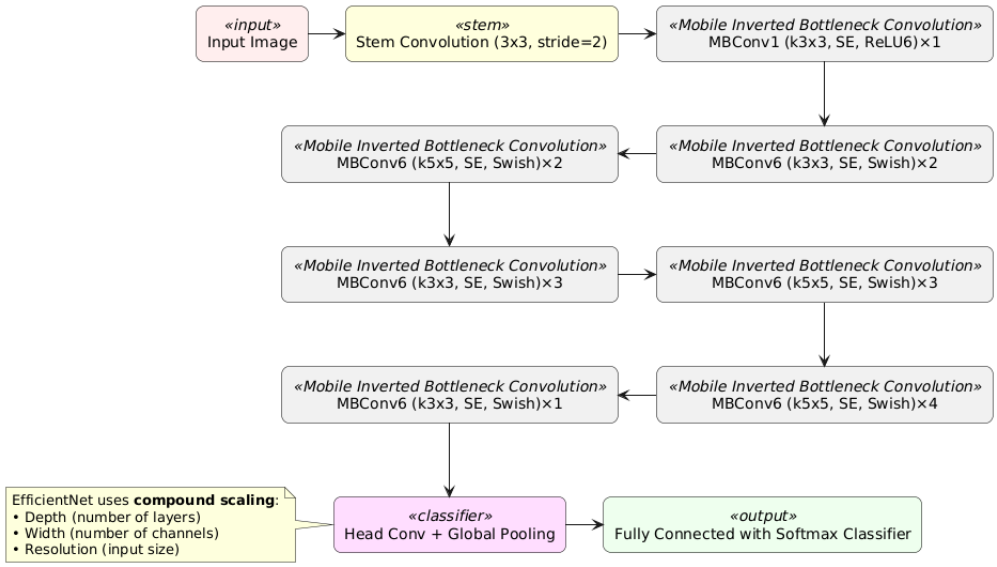


Fig. 2. The EffNet architecture (simplified)[71]

The EffNet, originally presented by Tan and Le in 2020 [71], represents a

significant shift in the CNN architecture by incorporating compound scaling, which simultaneously adjusts the network depth, width, and resolution. Initially optimized for image classification tasks on ImageNet, the EffNet series (ranging from B0 to B7) has reliably showcased enhanced accuracy-efficiency trade-offs. The unique compound scaling and architectural efficiency of this method have not only excelled in computer vision but have also motivated researchers to investigate its potential applications in non-visual fields, especially for processing spectrogram-based audio data.

Figure 2 depicts EffNet, which begins by processing an input image through a stem convolution layer. This layer serves to decrease spatial resolution while boosting the channel depth. Central to the design are mobile inverted bottleneck convolution (MBConv) blocks that stack to incorporate depth-wise separable convolutions, expansion layers, and squeeze and excitation (SE) mechanisms. These elements work together to capture local and global channel dependencies efficiently. The MBConv blocks vary in kernel size (either 3×3 or 5×5), activation functions (rectified linear unit (ReLU)6 or Swish), and their repetition count, achieving a balanced trade-off between representational capability and computational cost. Following the series of MBConv layers, a concluding head convolution and a global pooling layer synthesize the spatial data before submitting it to a fully connected softmax classifier for prediction. A unique feature of EffNet is the compound scaling, which systematically adjusts the depth, width, and input resolution of the network simultaneously. This approach improves the balance between accuracy and computational efficiency, surpassing models that focus on scaling a single dimension [71, 74]. Due to its capability and adaptability, EfficientNet has been used effectively in spectrogram-based audio classification tasks, including those in noisy anomaly detection environments [4, 86].

In 2021, the initial surge of audio-focused adaptations of EffNet centered on utilizing transfer learning (TL) for environmental sound datasets, like UrbanSound8k and ESC-10 [72]. Tsalera et al. (2021) [93] systematically compared pretrained CNNs, including those rooted in EffNet, to models such as VGG-ish [94] and YAMNet [95]. Their findings revealed notable performance improvements, achieving accuracies exceeding 96%, by fine-tuning image-based CNNs on sound spectrograms.

By the beginning of 2023, researchers in audio tagging had customized the EffNet-B2 architecture for this specific application. A significant study demonstrated that an attention-enhanced EffNet-B2, which added a multi-head attention module to the standard CNN framework, achieved audio tagging performance on par with ResNet-50 on AudioSet benchmarks, while preserving a considerably smaller size [88].

In mid-2024, a significant advancement occurred within the audio-centered domain CNN, although it was not directly associated with the exploration of EffNet. In their research, Choudhary et al. [96] introduced light and efficient audio classification network (LEAN), a hybrid model that underscores a lightweight

framework specifically designed for edge applications. This model integrates waveform encoders with logarithmic spectrogram backbones similar to those of YAMNet. While not directly founded on EffNet, this study emphasized the prevalent shift toward efficiency-oriented audio classifiers, mirroring the foundational tenets of EffNet's streamlined architecture.

In 2025, specialized adaptations of EffNet achieved significant breakthroughs. A study by He and Luo [97], published in Scientific Reports, introduced a customized EffNet-B0 model, optimized with efficient channel attention (ECA) and convolutional block attention module (CBAM) attention mechanisms, for bird song identification. This efficient model achieved an impressive accuracy of 96.04% among ten bird species, while reducing the parameter count by more than 16%, which highlights considerable efficiency improvements for practical biodiversity monitoring tools. By mid-2025, Porwal [89] employed a combination of EffNet-B0 and -B1 models on mel-spectrograms to classify bird species, and finished 25th out of 975 entries in the BirdCLEF 2024 challenge. This result demonstrates the flexibility of EffNet variants in handling demanding, long-duration audio-classification tasks.

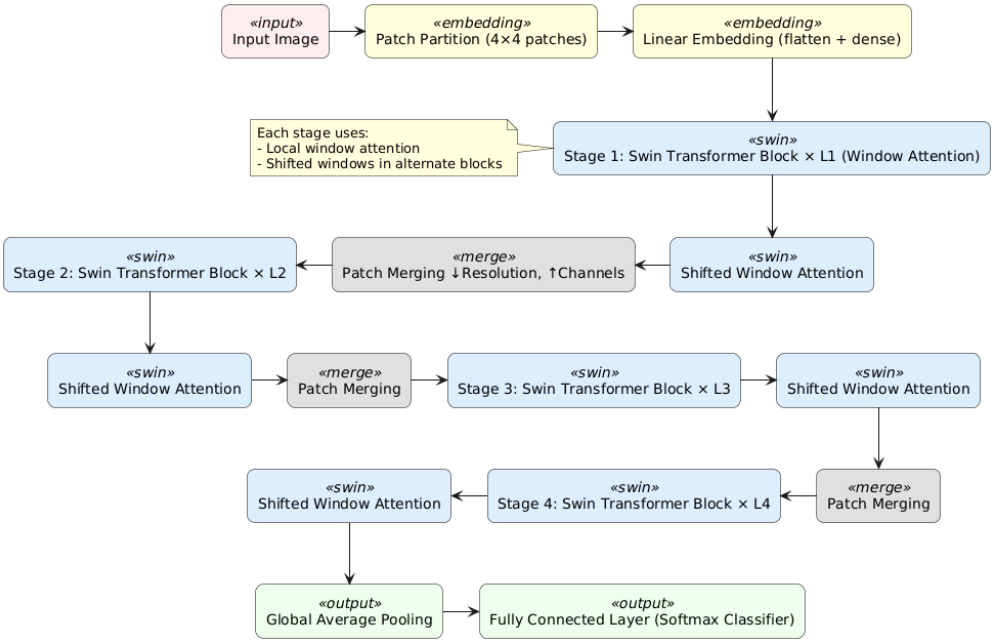


Fig. 3. The SWinT architecture (Simplified)[76]

This study adopts the SWinT as the primary transformer-based architecture for audio classification tasks. Introduced by Liu et al. in 2021, this model re-engineered the transformer (originally devised for NLP) for computer vision by incorporating a new shifted-window (SWin) mechanism. This design enhanced scalability and computational efficiency for image classification [76]. Liu et al. later advanced this

work in 2022 with SWinT V2, a revised version offering greater model capacity and improved high-resolution processing [77]. Unlike architectures such as EffNet, which are based on CNN, SWinT is built from transformer blocks that employ strategically shifted attention windows as the core of its classification strategy, enabling more flexible receptive fields and superior modeling of the global context.

Summarized in Figure 3, the SWinT starts by splitting an `Input Image` into fixed-size patches through *Patch Partition*, converting them into vectors using a `Linear Embedding` layer. These vectors pass through four hierarchical stages with stacked `SWin Blocks` by using `Window Attention` to capture local dependencies. Alternation with `SWin Attention` between stages facilitates data sharing between regions, while reducing the complexity of global attention [76]. Resolution decreases, and channel dimensions increase through `Patch Merging`, creating multi-scale representations similar to those found in convolutional hierarchies. The model concludes with the `Global Average Pooling` and `Fully Connected Layer with Softmax Classifier` to finalize the predictions. Integrating local window attention, SWin, and patch merging renders SWinT efficient and effective for structured inputs, such as spectrograms, thus ensuring strong audio classification [76, 77].

Recent research has expanded the benefit of SWinT architectures to auditory processing, adapting generally utilized computer vision DL models to audio signal processing by converting audio inputs into spectrograms. Zhao et al. (2022) [92] introduced self-supervised pre-training with shifted-windows transformers (S3T), a system that involves mel-spectrograms for music genre recognition, achieving significant accuracy rates of over 81% on the GTZAN dataset [98, 99] despite having limited labeled data. In 2024, Duan [100] further refined this model by developing broadcast shifted-windows transformers (BST), which improves the capture of low-level spectrotemporal details, achieving outstanding accuracy levels as high as 99% on the GTZAN dataset. Furthermore, Wang et al. (2024) [101] created the "Speech Swin-Transformers", utilizing a hierarchical shifted-window approach to enhance emotion recognition in speech, outperforming traditional speech processing methods. Moreover, Ramesh et al. (2024) [102] unveiled shifted window transformer emotion network (SwinEmoNet), an innovative speech emotion recognition (SER) model using SWinT to harness diverse spectrogram features, effectively addressing linguistic and cultural diversity, and achieving impressive accuracy rates of 94.93% and 96.51% on the Emo-DB[103] and RAVDESS [104] datasets, respectively, an thus surpassing current transformer-based techniques.

The empirical successes observed in audio classification highlight the versatility of the SWinT's SWin mechanism, extending its applicability beyond visual domains. Through the conversion of audio signals into time-frequency representations, the shifted windows effectively capture the inherent locality and overlapping characteristics, which mirror the hierarchical structuring of sounds, ranging from phonemes and words in speech to spectral-temporal patterns in music. This

alignment enables robust feature extraction across diverse tasks such as music genre identification, vocal emotion recognition, and keyword detection, thus establishing SWinT as a robust foundation for addressing classification tasks in both audio and visual contexts.

1.2. Acoustic Monitoring in Noise: Application Areas, Noise Profiles, and Anomaly Definition

The utilization of acoustic methods in the context of environmental surveillance or the identification of irregularities, as illustrated in Figure 1b, boasts a substantial historical precedent [105, 106]. In general, abnormal conditions that occur at the measuring device or location can be detected by changes in the characteristics of the generated acoustic signal, such as frequencies and amplitude [107, 108]. The advantage of using the acoustic method compared to other methods is that the features of the acoustic signal can be extracted and used for deeper failure detection. In addition, Tagawa et al. (2021) [48] stated that capturing acoustic data that involves only a microphone and is non-destructive facilitates the identification process without disrupting the running system. Furthermore, Reubens et al. (2019) [109] acknowledged that acoustic methods are also applied to detect changes in the behavior of living creatures. Unfortunately, during collection, compression, and transmission, all collected signals and acquired images are unavoidably polluted by noise, resulting in distortion and loss of information.

Research indicates that noise negatively impacts the efficacy of signal processing activities. Consequently, the significance of signal denoising has increased in contemporary signal processing systems, including applications in image processing [110], speech recognition [111], and biomedical signal processing, which is critical for medical diagnostics [112]. According to Damaševičius et al. (2017) [113], noise in telecommunications impairs communication channels, leading to bandwidth reduction and a subsequent signal quality decline, marked by jitter and information loss. Furthermore, Picaut et al. (2020) [114] asserted that urban noise adversely affects the health of city inhabitants and intensifies noise pollution. Additionally, Kantova et al. (2021) [115] observed that noise poses challenges in numerous industrial applications and the field of construction engineering.

Industrial noise is characterized as the sound that is present in workplaces and businesses, resulting from manufacturing activities and the use of machinery, tools, or equipment [116]. Exposure to such noise has several effects, including a decreased lifespan of industrial equipment and an increased risk of industrial accidents. Similarly, structural vibration resembles noise in its potential to harm the integrity and functionality of structures. Such vibration can lead to a range of negative outcomes, such as the onset of structural fatigue failure [117], discomfort among users or passers-by [118], interference with sensitive instruments, among others [119]. In engineering, the foundational step crucial to implementing condition monitoring and fault diagnostics entails a thorough evaluation of vibration data. The objective of this analysis is to identify the most critical problem features, thus improving the accuracy

of subsequent diagnostics and assessments. Therefore, accurately analyzing noise within the recorded vibration signals is crucial for precisely evaluating the unit's malfunctioning.

Unlike the industrial landscape, residential and forest regions possess unique acoustic attributes. Residential soundscapes are largely dominated by noises linked to human activities and potentially pets [120–122]. This sound environment includes a variety of noises such as everyday conversations, children playing, the hum of lawn mowers, barking of dogs, the hum of traffic, and even the sounds associated with brushing teeth. On the other hand, wilderness environments primarily feature natural sounds not related to human activities [123]. These natural acoustic elements might include wolf howls, tiger growls, lion roars, bird chirps, and gusts of wind. As a result, the parameters for identifying anomalies in these surroundings diverge substantially from those applicable to industrial settings.

This study considers industrial, urban, and forest environments not for their functional similarity, but for their complementary acoustic properties and noise conditions. Together, they span a wide range of real-world soundscapes in which anomaly detection systems must operate. Industrial scenes are dominated by repetitive, structured sounds, where anomalies usually reflect mechanical malfunctions. Urban scenes add highly variable, human-driven sounds and complex, non-stationary noise. Forest scenes, instead, feature unstructured natural soundscapes with overlapping biological and environmental sources, where anomalies are rare and hard to isolate. Examining these three settings together enables assessment of the robustness and generalization of DNN-based anomaly detection methods across diverse, acoustically challenging conditions.

In this dissertation, anomalies are not treated as brief, isolated acoustic events, but rather as distinct acoustic categories defined within the context of each dataset. An anomaly is therefore understood as a class or event type whose acoustic characteristics deviate from those associated with normal operating or environmental conditions in a given domain. Rather than attempting to localize anomalous events precisely in time, the proposed approach assigns fixed-duration audio segments or complete recordings to predefined sound classes, some of which explicitly represent anomalous behavior.

Importantly, the interpretation of what constitutes an anomaly is dataset-dependent. In industrial sound datasets, anomalous classes typically correspond to mechanical faults or abnormal machine operating states. In urban and residential datasets, anomalies may reflect uncommon or disruptive acoustic events relative to typical background sounds. In natural and forest sound datasets, anomalous classes may represent irregular biological activity or atypical environmental phenomena. Under this formulation, anomaly detection is realized through supervised sound classification, and the quality of the classification results directly determines the reliability of anomaly identification. This framing enables a unified evaluation of anomaly-related behavior across heterogeneous acoustic domains, even under noisy and non-stationary conditions.

1.3. Evolution of Acoustic Anomaly Detection Research: A Review Perspective

The merging of acoustic methods with the progress in AI has garnered growing attention from researchers, particularly in their use for anomaly detection. The timeline of progress in this interdisciplinary area is detailed through multiple reviews (traditional literature review (TLR), systematic literature review (SLR), methodological literature review (MLR), narrative literature review (NLR), etc.) that together explain the development, usage, and effectiveness of these techniques in different fields. Table 1 offers a compiled overview of these important reviews.

Table 1. Recent literature reviews in acoustic-based detections

Author(s)	Method	Citations	Focus of study
Shaikh et al. (2021) [27]	TLR	44	Exploring the potential of acoustic signal analysis in facilitating machine fault diagnosis.
Qurthobi et al. (2022) [124]	SLR	52	State-of-the-art mechanical failure detection using acoustic methods.
Bhuiyan and Uddin (2023) [125]	TLR	116	DTL implementation of industrial machine failures with vibration and acoustic data as the main source of analysis.
Sharma et al. (2023) [126]	MLR	54	Evaluating how AI technologies are applied in bioacoustic monitoring to enhance wildlife conservation efforts.
Brockmann-Bauser (2025) [127]	TLR	38	Recent developments, challenges, and future directions in AI for patients with voice disorder.
Manikandan and Neethirajan (2025) [128]	SLR	121	Engagement of AI-driven bio-acoustic method in sensitising stress level and disease presence in poultry farming.
Sebastián-González and Pérez-Granados (2025) [129]	SLR	306	In-depth examination of the geographic variation in acoustic signaling among wildlife.
Shokouhmand et al. (2025) [130]	NLR	97	Involvement of AI to analyze breathing sounds to enhance the diagnosis and monitoring of respiratory conditions.

Shaikh et al. began an early and comprehensive exploration in 2021 by TLR [27]. This investigation highlighted the fundamental notion that any marked deviation from standard acoustic patterns could indicate the presence of mechanical anomalies or functional disruptions. When accurately captured and analyzed, these deviations can reliably serve as indicators of system malfunctions or abnormal operating conditions. This perspective provided a foundation for future research that has increasingly been

integrated AI to improve anomaly detection capabilities.

In 2022, Qurthobi et al. conducted a SLR [124], analyzing 52 scholarly publications published between 2005 and 2021. Their comprehensive review encompassed a wide range of high-impact journals and conference proceedings across various disciplines. The authors concluded that the integration of acoustic methodologies with AI and DL significantly improves the detection and classification of mechanical damage, including, but not limited to, cracks, corrosion, surface wear, and pitting. The findings underscored that AI-enhanced acoustic analysis is both feasible and beneficial for early fault detection and predictive maintenance.

Building upon this momentum, Bhuiyan and Uddin in 2023 presented a comprehensive study TLR [125] that systematically evaluated the utilization of deep transfer learning (DTL) in the context of industrial fault diagnosis. Their review underscored the potential of DTL to address one of the most significant challenges in machine learning, namely, the scarcity of labeled data. The authors demonstrated that DTL enables the training of high-performance diagnostic models, even in scenarios with limited data availability, and showed particular efficacy in analyzing vibration and acoustic sensor signals. This contribution signifies a substantial advancement in enhancing the accessibility and scalability of AI-based diagnostics in practical industrial environments.

In 2023, Sharma et al. conducted a comprehensive MLR [126], redirecting attention toward bioacoustic applications within the framework of ecological conservation. The study conducted a critical analysis of various AI and ML methodologies for the processing and interpretation of wildlife acoustic data. The review highlighted significant enhancements in the accuracy and efficiency of species identification and behavioral pattern recognition, primarily driven by recent advances in AI and DL techniques. Besides these progressions, the research also illuminated several enduring challenges in the domain, including data heterogeneity, algorithmic biases, and the lack of standardized analytical protocols. The authors advocated for future research endeavors to address these limitations to enhance the robustness and reproducibility of ecological acoustic monitoring.

The year 2025 witnessed a notable surge in scholarly reviews that evaluated the current state and potential advances of acoustic methods, together with AI, ML, and DL, in various application domains. In particular, a significant TLR was undertaken by Brockmann-Bauser [127], who investigated the efficacy of AI, particularly DL architectures within the medical sector, with a focus on the detection and classification of voice disorders. The review demonstrated that well-structured DL models possess the ability to achieve high diagnostic accuracy, providing clinicians with improved tools to evaluate vocal pathology and improve the overall patient care quality through precise assessments based on acoustics.

In the domain of agriculture, Manikandan and Neethirajan conducted a thorough SLR [128] evaluation of the application of AI technologies in the evaluation of poultry vocalizations indicative of stress and disease. Their research highlights the

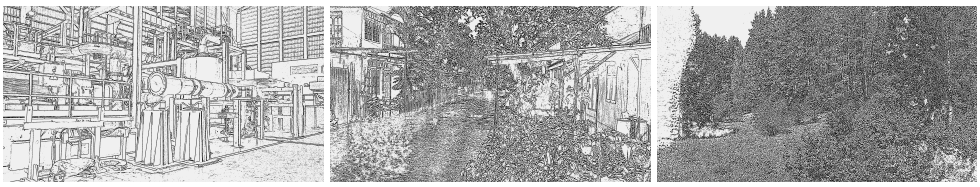
effectiveness of edge-AI and tiny machine learning (TinyML) solutions for real-time, non-invasive health monitoring, with significant implications for improving both animal welfare and agricultural productivity. These lightweight AI implementations are particularly suited for deployment in resource-constrained settings, thus facilitating scalable and cost-effective agricultural diagnostics.

In parallel, Sebastián-González and Pérez-Granados (2025) conducted an extensive SLR [129] analysis of 306 peer-reviewed studies examining geographic variation in wildlife acoustic signals. Their meta-analysis demonstrated that acoustic signatures tend to grow more distinct with increased geographic separation, a phenomenon most notably observed in avian species renowned for vocal learning. Nonetheless, the investigation also uncovered a substantial bias in taxonomic and geographic representation within current research, with songbirds being overrepresented and other taxa underexplored. To mitigate these imbalances, the authors called for more inclusive and diverse research endeavors across different species and regions.

In a NLR conducted by Shokouhmand et al. (2025) [130], the relatively unexplored domain of AI-driven analysis of nasal and oral breathing sounds in the context of respiratory disease diagnosis was examined. Although traditional chest auscultation is a well-established diagnostic method, the review highlighted that the potential of nasal and oral breath acoustics remains underutilized for clinical evaluation. The authors suggested that the implementation of AI techniques on these alternative sound sources could provide additional diagnostic insights, thereby assisting in the advancement of non-invasive and accessible respiratory diagnostic methodologies.

These reviews collectively underscore an emerging consensus on the transformative potential of integrating acoustic methodologies with advanced technologies of AI. In diverse domains, including industrial fault detection, medical diagnostics, wildlife conservation, and precision agriculture, the convergence of sound-based data with intelligent algorithms is constantly expanding the horizons of research and practical applications.

1.4. State of the Art Research in Audio-based Anomaly Detection



(a) industrial settings **(b)** human settlement **(c)** wilderness environments

Fig. 4. Portrayal of three distinct environments

This section presents the current state of research on classification based on

audio signals. Figure 4 shows the three types of environments (i.e., industrial areas, residential areas, and wilderness) that are investigated in this study.

1.4.1. Industrial Settings

In the realm of industrial machine condition monitoring and acoustic diagnostics, as it is illustrated in Figure 4a, the classification of audio signals is fundamentally grounded in the analysis of machinery-specific frequency patterns [26]. This methodological approach is based on the assumption that each type of industrial equipment (e.g., fans, pumps, sliders, and valves)[4] exhibits a unique acoustic signature that reflects its operational dynamics, such as internal vibrations, rotational frequencies, and thermal states. These physical phenomena manifest themselves as consistent auditory patterns during normal functioning, which can serve as diagnostic indicators. Consequently, deviations from these established patterns can serve as early warnings of mechanical degradation or failure, enabling the implementation of proactive maintenance strategies.

Traditionally, skilled human operators have been able to detect such deviations through auditory perception, particularly in acoustically controlled environments. In settings devoid of ambient interference, practitioners can often distinguish between nominal and anomalous machinery states without the need for specialized instrumentation. However, real-world industrial environments, as visualized in Figure 4a, rarely provide such ideal conditions. These settings are frequently saturated with continuous background noise generated by adjacent machinery, ventilation systems, or human activity, which collectively obscures critical audio signals [4].

To address these limitations, Purohit et al. [4] has introduced the MIMII dataset. This dataset comprises an extensive library of audio recordings from various industrial auxiliary machines (including pumps, valves, fans, and slide rails) captured under both normal and anomalous operating conditions. In an effort to emulate the acoustic complexity of industrial soundscapes, the clean audio recordings were deliberately augmented with synthetic environmental noise at three discrete SNR: -6, 0, and +6 dBs. This augmentation strategy was intended to promote model robustness and ensure the generalizability of classification algorithms in noisy settings.

Since its release, MIMII has become a cornerstone in the field of audio-based fault detection. It has catalyzed the development of a broad range of classification frameworks, particularly those leveraging DL architectures. Numerous studies have explored the application of ML and DL like the GAN [48], CNN [21], RNN such as LSTM [57], GRU, and more advanced pre-trained models [11, 133] to interpret and classify the complex acoustic features within MIMII recordings. These techniques often rely on advanced signal processing methods, including mel-spectrogram, MFCC, and other perceptually inspired representations, to extract salient characteristics from raw audio inputs.

Despite advancements in automated classification systems, human performance remains a relevant point of comparison. Research by Sakthivel et al. [135] highlights

Table 2. Summary of selected research in industrial machine condition monitoring

Author(s)	Method	Key Features & Findings	Acheivements
Koizumi et al. (2019) [131]	AE	Proposed unsupervised anomaly detection using AE on DCASE trained only on normal data; baseline for DCASE 2020 Task 2	89.2%
Koizumi et al. (2019) [39]	–	Developed a new dataset, ToyADMOS, using the recordings of miniature industrial machineries	–
Purohit et al. (2019) [4]	–	Released MIMII, a dataset of four machine types under normal/anomalous conditions; added synthetic noise at -6, 0, +6 dB SNR	–
Tagawa et al. (2021) [48]	GAN, SVM	Introducing a novel approach in MIMII classification by utilizing OC-SVM	99.79% (AUC)
Tanabe et al. (2021) [132]	–	Presented a MIMII-DUE, which is an update of MIMII by introducing a domain shift due to changes in environmental and operational settings	–
Thoidis et al. (2021) [133]	Deep-CNN	Used TL with deep-CNN; improved generalization to unseen anomalies on MIMII dataset	91.0% (AUC)
Ahn (2021) [21]	CNN	Applied CNN to log-Mel features for anomaly detection on MIMII; improved over baseline methods	$\sim 90\%$
Mobtahej et al. (2022) [57]	LSTM, GRU, AE	Implemented RNN-based models to capture temporal structure in machine sounds of MIMII; robust in noisy conditions	$\sim 100\%$
Guan et al. (2023) [11]	Transformer-based AE	Improved model generalization with contrastive learning for MIMII classification; better robustness across varying SNR	94.6%
Zaman et al. (2023) [134]	AI-enabled monitoring	Proposed remote real-time diagnostic system using DL for classifying MIMII; effective in hazardous settings	92.8%
Zabin et al. (2024) [23]	lightweight self-attention SqueezeNet	Introduced the combination of model with EMD-GS filter to classify MIMII and ToyADMOS	89.32% (MIMII) & 96.46% (ToyADMOS)
Zabin et al. (2025) [46]	Few Shots Learning-Based	Added few-shot learning-based to classify MIMII and ToyADMOS	89.6% (MIMII) & 98.3% (ToyADMOS)

the inherent limitations of manual monitoring in industrial settings, where persistent ambient noise complicates auditory assessments. In highly cluttered acoustic environments, human operators may experience auditory masking or fatigue, thereby reducing the reliability of manual fault identification.

This concern is further supported by recent empirical evidence from Toker et al. (2025) [136], who demonstrated that occupational exposure to elevated noise levels (specifically those surpassing 80.1 dB) adversely impacts worker concentration and task performance. These findings underscore the necessity for automated monitoring tools that can maintain high diagnostic precision irrespective of background interference.

In response to these challenges, Zaman et al. [134] have argued for the integration of AI- and DL-based methods into industrial diagnostic systems. Their research emphasizes that such technologies not only outperform traditional heuristic approaches in noisy environments but also reduce the dependency on human interpretation. Moreover, AI-driven solutions have proven instrumental in mitigating risks associated with hazardous operational conditions by enabling remote, real-time, and scalable acoustic monitoring frameworks.

To conclude, the development of industrial acoustic diagnostics has shifted from traditional, human-reliant monitoring to the use of advanced DL methodologies, propelled by the escalating complexity and noise levels prevalent in modern factory settings. The availability of reference datasets such as MIMII, MIMII-DUE, DCASE, ToyADMOS, alongside advanced neural network architectures and noise-resistant signal processing methods, has markedly improved the precision of early mechanical failure detection. Additionally, studies highlighting human limitations in noisy environments have further emphasized the necessity for automated solutions that offer reliable and scalable monitoring capabilities. These collective advancements highlight the increasing significance of incorporating AI into condition monitoring systems. Table 2 summarizes key studies and methodologies pertinent to this area.

1.4.2. Anthropogenic Habitats

In the domain of ESC, the ESC-50 dataset, presented by Piczak in 2015 [5], has established itself as a crucial benchmark for assessing ML and DL models focused on non-speech acoustic event recognition. It consists of 2,000 audio recordings, each lasting five seconds, and is evenly divided into 50 semantically balanced categories encompassing animal sounds, human activities, natural environments, and mechanical noises, as depicted in Figure 4b. Each audio clip was meticulously extracted and validated from user-contributed content in the Freesound.org database [37]. Owing to its well-organized structure and wide-ranging applicability, ESC-50 remains pivotal in the progression and evaluation of computational auditory scene analysis systems. Figure 4b offers a visual representation of the various environments that typically produce the audio categories included in the ESC-50 dataset. These environments range from residential and urban areas to public settings, where various environmental

and human-generated sounds are naturally present. By showcasing the common sound sources and settings covered by ESC-50, the figure helps to contextualize the diversity and relevance of the dataset in the real world. This visual insight complements the following paragraph, which delves into the structure, purpose, and significance of the dataset in the expansive field of ESC.

Table 3. Summary of selected research in anthropogenic habitat

Author(s)	Method	Key Features & Findings	Accuracy
Piczak (2015) [5]	-	Introduced ESC-50 with 2,000 curated 5-second samples across 50 balanced classes; established a standard environmental sound classification benchmark	–
Kong et al. (2020) [86]	PANN	Leveraged pretrained CNNs on AudioSet to provide transferable audio features; enabled strong TL for ESC tasks	–
Lin et al. (2020) [137]	ParallelNet	Utilized multi-branch CNN architecture for multi-resolution spectral-temporal learning; outperformed traditional CNNs on ESC-50	81.55%
Zhou & Zhao (2022) [138]	Transformer model	Applied self-attention to capture long-range temporal dependencies; demonstrated transformer effectiveness for ESC-50 classification	84.2%
Li et al. (2022) [139]	Attention-CNN	Enhanced time-frequency features with attention; achieved high accuracy across ESC-50, ESC-10, and UrbanSound8k	93.1%
Liu et al. (2023) [91]	CAT	Combined causal convolutions with unidirectional attention for real-time inference; improved to 97.2% with TL from PANN	96.9%
Gong et al. (2023) [140]	ACM-SSKD	Integrated self-supervised learning with knowledge distillation; achieved SOTA accuracy via hybrid training approach	98.7%
Chen et al. (2025) [69]	MobileNetV2 + SPA	Introducing extension for ESC-50 by adding noise sample from quadrotor UAV; developed lightweight CNN using spectral pooling and attention; maintained high performance with lower complexity	91.75%
Bouaziz et al. (2025) [141]	AST + data poisoning	Investigated adversarial robustness by using data poisoning	$\sim$95%

Since its formation, ESC-50 has established itself as a leading standard in the field of environmental sound analysis. In 2020, Kong et al. [86] introduced pre-trained audio neural networks (PANN), trained on expansive audio datasets such as AudioSet [17]. These models showcased significant transfer capabilities

to various downstream applications, including ESC-50 classification, by extracting comprehensive, versatile audio representations. Lin et al.'s 2020 study [137] presented parallel network (ParallelNet), an innovative DL architecture specifically designed for ESC-50. This model employs multiple parallel convolutional branches to capture spectral-temporal information at multiple resolutions. Their findings achieved an accuracy of 81.55%, marking a significant advancement over previous standards.

Various methods have been proposed to address the ESC-50 challenge. In 2022, researchers sought to resolve classification challenges by using an transformer-based classification network. Zhou and Zhao [138] investigated transformer models for ESC-50, which, through self-attention mechanisms, adeptly captured extended dependencies in audio spectrograms, thereby achieving 84.2% classification accuracy. Later, Li et al. [139] integrated an attention-enhanced CNN to optimize feature weighting in time-frequency representations. Their approach achieved 93.1% accuracy on ESC-50 and exhibited exceptional performance on ESC-10, a subset of ESC-50, as well as the UrbanSound8k dataset.

A subsequent pioneering method emerged to improve ESC-50 classification accuracy. In 2023, Liu et al. [91] presented the CAT, which utilized causal convolutions and unidirectional attention to enable real-time inference emulation. Their model attained a 96.9% accuracy on ESC-50. When complemented by TL with PANN [86], the accuracy increased to 97.2%. Also in 2023, Gong et al. [140] introduced the audio classification method (ACM)-self-supervision with knowledge distillation (SSKD) framework, combining self-supervised learning with knowledge distillation. Their approach reached a remarkable classification accuracy of 98.7% on ESC-50, highlighting the success of hybrid training strategies.

In recent times, scholars have integrated optimization and unconventional strategies to diversify the classification suite. In 2025, Chen et al. [69] introduced a lightweight model anchored in MobileNetV2 [73], enhanced by spectral pooling attention (SPA) mechanisms. This design attained a performance rate of 91.75%, demonstrating that efficiency-focused models can still produce impressive outcomes. Meanwhile, Bouaziz et al. (2025) [141] investigated an atypical methodology by implementing data poisoning techniques together with audio spectrogram transformers (AST). While precise figures were not numerically published, visual representations suggested a classification accuracy near 95%.

As outlined in Table 3, these significant research efforts illustrate the advancing complexity of ESC. Scholars have systematically evolved from simple CNNs to more advanced attention-based and transformer frameworks, frequently augmented with pretraining and self-supervised methodologies. The ESC-50 dataset has played an essential role as a benchmark for these advancements, facilitating comparative evaluations and encouraging the creation of more resilient sound classification algorithms.

1.4.3. Wilderness Environment

The soundscape within the cores of wilderness regions presents an extraordinarily complex and dynamic array of sounds, greatly exceeding the simpler soundscapes previously examined. This intricacy is due to the rich interaction of various biotic and abiotic elements [53, 123, 142]. In contrast to the predictable sound patterns typical of urban or industrial areas, dominated by human noise, the sonic imprints of wilderness spaces are characterized by significant variability. The collaborative effects of seasonal changes, diverse biological communities, impressive geological features, and fluctuating climatic conditions shape these soundscapes. Understanding and interpreting the auditory narratives of these natural settings is essential for precise ecological monitoring, the development of conservation strategies, and the progress of remote environmental sensing technologies. Through the analysis of these soundscapes, researchers can observe undisturbed wildlife, identify subtle signs of human interference, and assess the health of whole ecosystems. Thus, creating highly specialized datasets and classification systems that reflect the complex character of wilderness soundscapes is crucial to advancing the innovative field of environmental audio research.

The FSC22 represents a significant advancement in the field of environmental sound analysis, particularly in relation to ecological and conservation-related audio classification tasks. As introduced by Bandara et al. [7], this dataset was curated with meticulous attention to capture the diverse range of acoustic phenomena from wilderness ecosystems, including forests and remote natural habitats, as illustrated and visualized in Figure 4c. In contrast to more general sound datasets, such as ESC-50, which encompass a wide range of anthropogenic and natural audio events, FSC22 is specifically designed to support research efforts in bio-acoustic monitoring, wildlife surveillance, and eco-acoustic assessment.

The body of research on wilderness setting audio classification, which is notably involving the FSC22 dataset, highlights its growing importance as a key benchmark in forest sound classification, particularly in ecological audio analysis. Since its inception, FSC22 has fueled a variety of research efforts, ranging from traditional machine learning methods to sophisticated DL architectures. Designed with real-world wilderness audio recordings that encompass both biotic and abiotic sources, the dataset provides a challenging and realistic platform for model development. Collectively, all these studies depict a maturing research field where both architectural and dataset innovations are valued as complementary routes to create precise, efficient, and ecologically sound environmental sound classification systems. A comprehensive summary of these contributions and their performance metrics is available in Table 4.

The specificity of the sound events represented in FSC22, including animal vocalizations, abiotic environmental sounds, and various forms of forest activity, has stimulated the development of novel algorithmic strategies tailored to the challenges inherent in ecological audio classification. The complexity of the dataset, which

Table 4. Summary of selected recent research on audio-based classification for wilderness datasets

Author(s)	Dataset	Model/Method	Key Features & Findings	Accuracy
Bandara et al. (2023) [7]	FSC22	Dataset proposal	Introduced FSC22 for ecological sound classification; includes biotic, abiotic, and forest-specific acoustic events	92.16%
Ranmal et al. (2024) [143]	FSC22	ESC-NAS	Neural architecture search optimized model for edge deployment; significantly boosted classification accuracy	85.79%
Xu & Chen (2024) [144]	FSC22	ERT	ERT outperformed SVM, kNN, and DTs; ensemble-based methods suitable for complex wilderness audio	66.00%
Paranayapa et al. (2025) [145]	FSC22	CNNs + Compression	Applied pruning, quantization, and augmentation; DenseNet-121 achieved best results, but the utilization of MobileNet-v3 and ACDNet balanced the computational load and accuracy	99.22%
Ahmad et al. (2024) [146]	EFSC-24	kNN, SVM, RF, XGBoost	Enhanced the FSC22 by adding more transcripts	$\sim$ 89.87%
Chasmai et al. (2025) [147]	iNaturalist	Dataset proposal	Collecting enormous audio transcripts collection from diverse animals	–
Qurthobi et al. (2025) [148]	FSC22	CNN + BiLSTM	GoogLeNet with BiLSTM using MFCC features	78.52%

originates from overlapping sources, non-stationary noise, and the spectral similarity of various sound categories, presents significant hurdles for traditional machine learning models. In response to these difficulties, researchers have explored a range of approaches aimed at maximizing classification performance.

Only one year after its inception, more researchers became interested in examining this dataset. Xu and Chen [144] conducted a comparative analysis of conventional machine learning algorithms and reported that the use of the extremely randomized trees (ERT) method yielded the highest classification accuracy among the models tested. Their method achieved an accuracy score of 66%, surpassing SVM, k-nearest neighborhood (kNN), and decision tree (DT)-based models. The efficacy

of ensemble-based approaches in this context suggests their potential to model complex acoustic patterns, especially when data availability is constrained. Ranmal et al. (2024) [143] adopted a more satisfactory approach through the deployment of environmental sound classification using hardware-aware neural architecture search (ESC-NAS), a framework designed to optimize neural architectures under computational constraints. The application of ESC-NAS resulted in a substantial performance improvement, achieving an accuracy of 85.79% on the FSC22 dataset. This outcome highlights the advantage of automated architecture search in identifying optimal configurations for environmental sound classification tasks, particularly when deployed on edge devices with limited resources.

Paranayapa et al. (2024) [145] carried out a broader study in forest sound classification, focusing on how effectively model compression techniques and audio preprocessing approaches could enhance DL-based ESC. This research utilized the FSC22 dataset [7], a key resource for forest acoustic monitoring. Their work included a thorough comparison of various CNN-based models, ranging from simple, lightweight structures to more sophisticated ones like adaptively combined dilated convolution for monocular panorama depth estimation (ACDNet) [149], AlexNet, and ResNet-50, all within an optimized processing framework. This framework employed advanced preprocessing techniques such as time-stretching and pitch-shifting, along with various spectrogram representations, and iterative model compression methods, including weight and filter pruning, as well as 8-bit quantization. The research showed that integrating time-frequency augmentation with spectrogram fusion consistently improved model robustness and generalization. Among the studied models, DenseNet-121 stood out as the most accurate, achieving a classification accuracy of up to 99.22%. When considering accuracy, memory, and inference speed trade-offs, especially for edge deployment, MobileNet-v3-small and ACDNet provided the best balance, achieving accuracies of 87.95% and 85.64% respectively, even with strong compression.

Complementing these findings, Ahmad et al. (2024) [146] introduced the EFSC-24, an alternative strategy focused on dataset-level enhancement rather than architectural complexity. By enriching the original FSC22 dataset with additional, meticulously labeled audio samples representative of various forest acoustic events, their study sought to improve model performance through data diversity and volume expansion. Notably, even without employing DL techniques, their approach resulted in a substantial performance gain by using the XGBoost algorithm, a tree-based ensemble method known for its computational efficiency and interpretability. The model achieved a notable classification accuracy of 89.87%, highlighting the potential of data-centric approaches in enhancing classification outcomes. This result underscores the broader principle that, while advanced architectures and optimization pipelines can significantly improve model performance, enhancements in training data quality and quantity can also produce comparable benefits, especially when paired with well-calibrated traditional machine learning models. Collectively, these studies

affirm the viability of both model-centric and data-centric strategies for improving environmental audio classification systems, particularly in applications where computational resources and deployment constraints must be carefully managed.

More recently, Qurthobi et al. (2025) [148] introduced a hybrid DL architecture that integrates CNN with BiLSTM. Their model, which combines the spatial feature extraction capabilities of GoogLeNet with the temporal modeling strength of BiLSTM, demonstrated promising results by achieving an accuracy of 78.52%. The model used MFCC as the primary characteristic representation, using its perceptually motivated characteristics to capture both timbral and temporal dynamics in the forest soundscapes. Although the proposed method did not exceed the ESC-NAS framework in raw accuracy, it provides a viable balance between the complexity, interpretability, and performance of the model.

1.5. Summary and Motivation for the Present Study

1.5.1. Synthesis of Prior Work and Open Challenges

Historically, acoustic signals have played a vital role in spotting anomalies across various fields, such as mechanical diagnostics, structural health monitoring, and environmental surveillance [105, 106]. Originally, the detection of acoustic anomalies relied on human auditory skills, with operators using their hearing to notice deviations in sound, like frequency changes, unusual resonances, or irregular vibrations from machinery or environmental setups. This method was founded on the widely accepted notion that changes in a system's acoustic properties usually signal potential issues or failures [107, 108]. Although practical and requiring minimal gear, this approach suffered from inherent subjectivity and variability. The results were highly contingent on the individual's expertise, hearing ability, and awareness of the context, leading to inconsistency and restricted reproducibility. Additionally, the human ear's limitation in detecting minor acoustic shifts often delays the identification of early-stage faults, raising the risk of operational disruptions or environmental harm. As systems grew more complex, needing constant real-time monitoring, the drawbacks of manual auditory assessments became more evident. This situation propelled the shift towards more quantitative and automated approaches using signal processing and intelligent algorithms to evaluate and interpret acoustic data [48], laying the groundwork for modern AAD systems.

As digital signal processing has advanced, the constraints associated with manual auditory detection have prompted the development of systematic approaches for analyzing acoustic data. Researchers began to extract measurable features from audio signals, including amplitude shifts, dominant frequencies, spectral entropy, and energy distribution. These features are essential for assessing a system's performance and condition. By using these signal-level descriptors, raw audio waveforms could be converted into structured data, which is more compatible with algorithmic analysis. This transformation led to frameworks that can identify anomalies by spotting deviations from anticipated or learned signal patterns, fostering a more objective and reproducible method for anomaly detection [48]. Unlike manual methods, these

systems offer improved sensitivity and time resolution, enabling the detection of faults at an early stage and reducing dependence on human interpretation. This shift marked the start of an era of scalable acoustic monitoring systems that support automated analysis across various domains and operational levels, from individual mechanical parts to large-scale environmental monitoring networks.

The integration of ML techniques further enhanced the capabilities of AAD by enabling systems to learn distinctive remarks from labeled data. The well-known ML classifiers, e.g., SVM, DT, kNN, etc., were among the earliest tools adopted for this purpose. These models allowed the construction of statistical and rule-based frameworks capable of distinguishing between normal and anomalous acoustic events [27, 124]. Their effectiveness was particularly evident in structured environments where the types of anomalies and failure modes were well defined and consistently represented in the training data. By mapping the features of the extracted signal to specific operational states, the algorithms ML facilitated the automation of fault recognition and reduced the dependency on manual inspection. However, these models often required extensive feature engineering and struggled with generalization in dynamic or noisy acoustic settings, thus setting the stage for more robust solutions enabled by DL.

The ongoing rise of DL signifies a groundbreaking shift in the identification of acoustic anomalies, offering unparalleled progress in both accuracy and flexibility. In the current field, DNNs, particularly structures based on CNNs and RNNs, are consistently utilized to thoroughly examine both unedited and preprocessed audio signals [48]. These advanced models are remarkably adept at autonomously learning complex, hierarchical, and abstract data representations, effectively capturing the essential temporal and spectral cues necessary for anomaly detection. In contrast to conventional machine learning techniques, DLs approaches significantly reduce reliance on manually designed features, facilitating superior generalization even in challenging and noisy settings. The resilience and scalability of these models have established them as fundamental components in modern audio detection systems, particularly in contexts that demand heightened sensitivity and real-time processing capabilities [106, 107].

The industrial sectors have experienced exceptional advantages due to these revolutionary innovations. Shaikh et al. (2021) [27] demonstrated the outstanding effectiveness of using DL in acoustic diagnostics, achieving unmatched precision in identifying mechanical faults compared to the limited accuracy of conventional methods. Qurthobi et al. (2022) [124] highlighted the transformational impact of real-time classification systems, employing NN for predictive maintenance and proactive early fault detection in rotating machinery with excellent timing and precision. Bhuiyan and Uddin (2023) [125] proposed an innovative multimodal strategy that combines acoustic and vibration data, significantly enhancing the dependability of fault diagnosis in complex and demanding mechanical environments. Additionally, in a step towards ecological synchronization, bioacoustics applications

have gained traction, as Sharma et al. (2023) [126] effectively used deep models for identifying species-specific sounds, greatly aiding wildlife monitoring and advancing conservation initiatives. Together, these pivotal studies underline the exceptional flexibility and power of acoustic signals as a vital tool for detecting key events across both industrial and environmental realms.

Despite these successes, several challenges persist in deploying acoustic-based anomaly detection systems in real-world settings. Environmental noise, data imbalance, and variability in operational contexts continue to hinder generalization and robustness [127]. Furthermore, limited access to large, annotated acoustic datasets remains a bottleneck for training deep models effectively. Nevertheless, recent advancements in low-power edge-computing devices, such as embedded digital signal processors and microphones, have made it feasible to implement real-time inference in constrained environments. Looking forward, a crucial research trajectory lies in optimizing DL architectures for anomaly detection under non-stationary conditions, enhancing model interpretability, and developing standardized benchmarks to support reproducible evaluation. As this field evolves, the synergy between acoustic sensing and intelligent algorithms promises to revolutionize event detection across diverse domains [105, 108].

1.5.2. Motivation and Design of the Comparative Study

In the realm of anomaly detection for noisy audio, it is imperative to properly select both the characteristic representation and the model's backbone to ensure effective outcomes. To mitigate the risk of becoming overly reliant on any singular feature, model, or dataset, we conducted an evaluation involving chroma-STFT, MFCC, and mel-spectrogram characterizations, alongside backbone options such as EffNet, SWinT, and their respective temporal variations. This assessment was performed across the datasets MIMII, ESC-50, and FSC22 as referenced in [4, 5, 7].

Mel-spectrogram, which maintains intricate spectral–temporal details, serves as a solid baseline for various sound events [5]. MFCC, on the other hand, provides a concise and noise-resistant summary of the spectral envelope, suitable for scenarios with limited resources [124]. Nevertheless, chroma-STFT highlights harmonic structures and ensures octave consistency, assisting tonal irregularities, although it may overlook broadband transients. Comparing these methods reveals whether the robustness of the noise is due to detail, compactness, or harmonic emphasis.

The backbones incorporate varying inductive biases. EffNet excels at efficient local feature extraction, which is ideal for low-latency edge scenarios. In contrast, SWinT effectively captures long-range dependencies and multiscale patterns, which are beneficial for industrial faults or prolonged calls from wildlife [76, 77]. Furthermore, temporal modules, such as BiGRU or BiLSTM, can enhance sequential modeling, especially for ecoacoustic data [7].

Building upon the identified challenges and objectives, this study is guided by three central research questions. The first concerns the effectiveness of different

acoustic features, namely, MFCC, mel-spectrogram, and chroma-STFT, in resisting noise and maintaining robustness under diverse SNR conditions, particularly in industrial datasets such as MIMII. The second question asks whether attention-based architectures, such as SWinT and its temporal variants, can provide superior cross-domain generalization compared to convolutional models when applied across different datasets, for example, from ESC-50 to FSC22. The third question focuses on how efficiency-oriented models can be adapted to balance accuracy, latency, and memory consumption, thereby ensuring practical applicability in embedded and resource-constrained systems [86, 124]. Together, these questions connect the methodological exploration with real-world deployment challenges, ensuring that both theoretical insights and practical constraints are systematically addressed.

2. DATASETS AND METHODOLOGIES

This chapter presents the materials and methods used to develop and evaluate the proposed framework for audio classification and anomaly detection. Section 2.1 describes the datasets used in this study, emphasizing their acoustic characteristics and relevance to industrial, urban, and natural environments. Section 2.2 introduces the time–frequency feature extractors for audio representation, summarizing their theoretical basis and practical role in sound classification. Section 2.3 outlines the proposed deep-learning architectures, detailing their design choices and integration strategy. Section 2.4 details the k-folds CV procedure used to obtain statistically reliable performance estimates. Section 2.5 summarizes the experimental workflow, including data processing, training, and evaluation. Finally, Section 2.6 describes the working environment, specifying the hardware and software platforms used in the experiments.

2.1. Datasets

To thoroughly evaluate audio classification systems, it is essential to test them across various acoustic environments. Thus, three well-known and publicly accessible datasets (i.e., ESC-50, MIMII, and FSC22) were chosen, each representing distinct environmental sound categories. The ESC-50 dataset comprises 2,000 audio clips, each lasting five seconds, categorized into 50 distinct semantic categories, including natural, human-created, and urban sound types. Due to its balanced array of classes and diversity, it has become a widely recognized benchmark for non-speech environmental sound identification [5]. Developed by Purohit et al., the MIMII dataset features audio recordings from industrial machines like valves, pumps, fans, and sliders, captured under both normal and anomalous situations. This dataset is extensively used in anomaly detection and predictive maintenance studies [4]. On the other hand, the FSC22 dataset, created by Bandara et al. [7], targets forest acoustic surveillance and includes 2,025 labeled audio fragments across 27 categories, covering biotic, anthropogenic, and geophysical sounds. This makes FSC22 especially pertinent for environmental monitoring, biodiversity research, and forest surveillance. Usage of these datasets provides an opportunity to validate classification techniques across general, industrial, and ecological acoustic landscapes.

The selection of the ESC-50, MIMII, and FSC22 datasets is motivated by their joint coverage of the main acoustic environments studied in contemporary audio classification. The ESC-50 dataset includes a broad set of everyday urban and domestic sounds, making it suitable to assess how well models generalize under varied and unconstrained conditions [5, 87]. In contrast, MIMII targets industrial soundscapes, where machine operations follow characteristic temporal patterns but also exhibit subtle deviations related to fault or wear that are central to anomaly detection and condition monitoring [4, 131]. Complementing these, FSC22 captures natural and semi-natural outdoor soundscapes, with an emphasis on forest environments that feature biotic activity, human presence, and geophysical

processes, which are key priorities in eco-acoustics and environmental monitoring [7, 53]. Together, these datasets support a comprehensive assessment of audio classifiers across general environmental, industrial, and ecological domains, enabling robust conclusions about model performance in diverse acoustic contexts common in sound analysis research.

2.1.1. MIMII

The MIMII dataset is a comprehensive collection of acoustic data designed to support research in the field of industrial machine condition monitoring and fault diagnosis. In 2019, Purohit and colleagues from the research and development (R&D) Group at Hitachi Ltd. compiled and publicly released the document. As documented in their report [4], the dataset captures sound recordings of standard industrial machinery operating under realistic factory conditions. Specifically, it encompasses audio samples from four distinct types of equipment commonly found in manufacturing environments: fans, pumps, sliders, and valves.

Table 5. Distribution of audio files in the MIMII dataset [4]

Device	ID	Condition		Device	ID	Condition	
		Normal	Abnormal			Normal	Abnormal
Fan	id_00	1011	407	Slider	id_00	1068	356
	id_02	1016	359		id_02	1068	267
	id_04	1033	348		id_04	534	178
	id_06	1015	361		id_06	534	89
Pump	id_00	1006	143	Valve	id_00	991	119
	id_02	1005	111		id_02	708	120
	id_04	702	100		id_04	1000	120
	id_06	1036	102		id_06	992	120

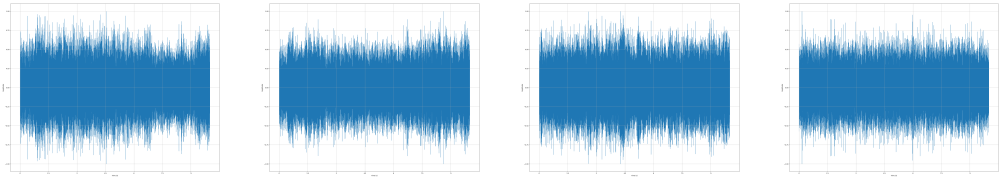
The MIMII dataset offers an in-depth overview of audio recordings from manufacturing settings, documenting the distinct operational conditions of various machinery in both normal and defective states. Table 5 outlines the distribution of recordings within this dataset, providing a foundation for a comprehensive analysis. Meanwhile, Table 6 and Figure 5 depict its general audio properties and selection of its audio transcriptions waveforms, respectively. This data set is vital for the development and rigorous testing of ML algorithms for improved anomaly detection. Implementing a robust data management strategy improves the efficiency of analysis processes. In this well-structured framework, audio recordings are meticulously categorized by the machine type, specific version with unique identifiers, and operational state. As such, the MIMII dataset is crucial for researchers endeavoring to enhance the accuracy and dependability of industrial diagnostic tools.

The MIMII dataset stands out as a key resource for advancing ML in the realm

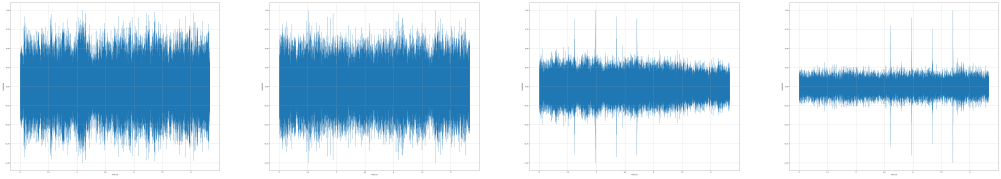
Table 6. General properties of the MIMII dataset

Property	Value
Total number of files	∼55,000+
File duration	10 seconds
Sampling rate	16 kHz
Channels	Stereo
Bit depth	16-bit
File format	WAV
Classes	2 (operations only)
	8 (2 (operations) × 4 (machines))
	32 (2 (operations) × 4 (machines) × 4 (identifiers))
Files per class	Varies

of industrial fault detection, providing a diverse and realistic collection of sound recordings from different types of machines operating in both normal and defective states. The detailed organization of these recordings by the machine category, identifier, and operational condition enhances the dataset's effectiveness in developing sturdy diagnostic systems. As indicated in Table 5, there is a noticeable imbalance in the number of samples between normal and faulty conditions across all machine variants, with significantly fewer examples of malfunctions. This imbalance presents challenges for model training and could potentially distort classification results. Therefore, it is essential for future research to address this issue by implementing strategies such as data augmentation, resampling, or cost-sensitive learning to ensure dependable and widely applicable performance in practical settings.



(a) normal fan **(b)** abnormal fan **(c)** normal pump **(d)** abnormal pump



(e) normal slider **(f)** abnormal slider **(g)** normal valve **(h)** abnormal valve

Fig. 5. Normalized waveforms of various audio samples in the MIMII dataset[4]

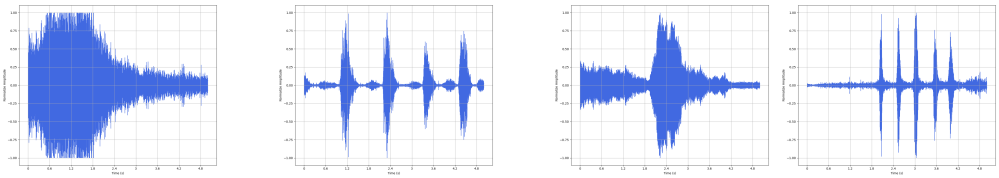
In this work, individual machines were not modeled as separate classes. Instead, audio samples were aggregated by device category and operating condition, yielding a

streamlined class hierarchy that enhances model generalization and limits overfitting to particular units. This approach is consistent with recent industrial acoustic monitoring studies and enables reliable anomaly detection in noisy settings.

2.1.2. ESC-50

Table 7. Distribution of the ESC-50 dataset [5]

Categories	Contents
Animals	cat, cow, crow, dog, frog, hen, insect, pig, rooster, and sheep.
Natural Soundscapes	chirping birds, cracking fires, crickets, pouring water, rain, thunderstorm, toilet flush, sea waves, water drops, and wind.
Non-speech	breathing, brushing teeth, clapping, coughing, crying baby, drinking/sipping, footsteps, laughing, sneezing, and snoring.
Indoor/Domestic	can opening, clock alarm, clock tick, door knock, door/wood creaks, glass breaking, keyboard typing, mouse click, vacuum cleaner, and washing machine.
Outdoor/Urban	airplane, car horn, chainsaw, church bells, engine, fireworks, handsaw, helicopter, siren, and train.



(a) airplane **(b)** breathing **(c)** cat **(d)** dog

Fig. 6. Normalized waveforms of various audio samples in the ESC-50 [5]

The ESC-50 dataset represents an important standard reference in the realm of environmental audio data collections, carefully designed to encompass a wide array of sounds commonly encountered in both residential and urban settings. This dataset, introduced by Karol J. Piczak in 2015 [5], was crafted with attention to detail through a manual selection process, utilizing audio samples sourced from the collaborative community platform Freesound.org [37]. One of the defining attributes that separates the ESC-50 dataset from other collections, such as MIMII, is its equitable partition of sound samples. The dataset includes a total of 2,000 distinct five-second audio clips, equitably distributed across 50 clearly defined semantic categories, providing 40 examples per category. This uniform distribution plays a crucial role in promoting consistent and fair conditions for the training and evaluation of ML and DL models focused on the recognition of environmental sounds.

To ensure consistent evaluation and reproducibility, the ESC-50 dataset includes a predefined k-folds CV framework, with $k = 5$ [5]. This division guarantees that each

fold preserves class equilibrium and includes representative samples from the full spectrum of sound categories [150]. This setup not only enables fair benchmarking among different models, but also reduces the risk of overfitting by systematically rotating the training and validation sets. This methodological uniformity has made ESC-50 especially qualified for comparing various ML and DL techniques under standardized conditions.

Table 8. General properties of ESC-50 dataset

Property	Value
Total number of files	2,000
File duration	5 seconds
Sampling rate	44.1 kHz
Bit depth	16-bit
Channels	Stereo
File format	WAV
Classes	50
Files per class	40

Piczak, the sole creator of ESC-50, improved the organization of the 50 sound classes by sorting them into five principal categories: animal sounds (e.g., dog barking and rooster crowing); natural ambient sounds (e.g., rain and thunderstorms); human activities (e.g., coughing and sneezing); indoor sound events (e.g., clock ticks and washing machines); and outdoor environmental sounds (e.g., car horns and chainsaws). Table 7 and Figure 6 reveal the structure and samples of the audio waveforms in this dataset. Furthermore, Table 8 shows the general properties of its transcriptions. This hierarchical structure facilitates the execution of more complex analyses and enables tasks such as hierarchical classification and transfer learning. The design aims to improve the interpretability and organization of the dataset. As a result, the ESC-50 dataset has become a widely recognized benchmark in the field of environmental audio classification, particularly for evaluating the performance of DL models, due to its carefully curated content, balanced class distribution, and relevance to real-world acoustic environments. Studies by Nanni et al. (2021) [151] and Zhang et al. (2019) [152] have validated ESC-50 as a reliable and comprehensive benchmark for environmental sound classification research.

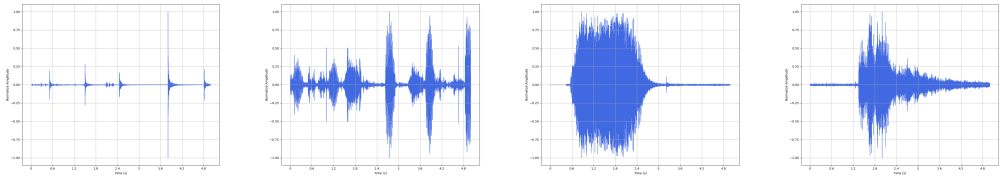
2.1.3. FSC22

The year 2023 marked the introduction of the FSC22 dataset, representing a noteworthy stride in the ever-progressing domain of environmental sound classification. Unlike prior datasets such as ESC-50, this groundbreaking collection was meticulously crafted to capture the raw auditory landscapes of pristine wilderness regions. It offers an innovative angle through its well-defined goals and structured design. The term "22" in its name refers to the year the dataset was inaugurated, and not to the number of classes, as in the ESC-50. Featuring a remarkable assortment of

unblemished audio recordings, the FSC22 encapsulates the complex harmony of both biotic and abiotic sounds, thereby providing a vibrant auditory depiction of the richly varied forest ecosystems.

Table 9. Taxonomy of the FSC22 dataset [7]

Threatening	Hazard	fire
	Intruders	speaking, whistling, clapping, footsteps, vehicle engine, and helicopter
	Tree-logging	axe, chainsaw, handsaw, generator, tree-falling, and wood chopping
	Poaching	gun-shot, and fireworks
Non-threatening	Animal	squirrel, lion, wolf, bird-chirping, bird-wing-flapping, frog, and insects
	Weather	thunderstorm, wind, rain, and waterdrops
	Silence	silence



(a) axe **(b)** bird-chirping **(c)** speaking **(d)** thunderstorm

Fig. 7. Normalized waveforms of various audio samples in the FSC22 [7]

Table 10. General properties of the FSC22 dataset

Property	Value
Total number of files	2,025
File duration	5 seconds
Sampling rate	44.1 kHz
Bit depth	16-bit
Channels	Stereo
Classes	27
Files per class	75

As indicated in the publication by Bandara et al. (2023) [7], the FSC22 was curated to fulfill an urgent demand for a diverse and extensive audio collection capturing the complexity of authentic forest ecosystems. It comprises 2,025 annotated audio clips, each of which is assigned to one of 27 unique sound categories. These categories hold contextual importance and are divided into threatening and non-threatening auditory phenomena, aiding their application in automated monitoring and ecological observation efforts. This classification is crucial for

improving the performance of ML models that aim to identify and interpret sounds that could indicate changes or hazards in forest settings [53].

The FSC22 holds particular relevance for domains such as eco-acoustics, biodiversity monitoring, and the development of autonomous threat detection systems. Its contribution extends to practical conservation efforts by enabling near real-time identification of anomalous sounds such as gunshots, chainsaws, or various animal calls [148, 153]. The inclusion of a wide array of natural and anthropogenic sound events ensures the dataset's adaptability across various research objectives. The structural composition of the dataset is illustrated in Table 9, while Figure 7 showcases the waveform diversity across representative classes, emphasizing the acoustic variability that makes FSC22 an indispensable benchmark for future research in environmental sound recognition.

2.2. Feature Extractors

To ensure strong and precise audio classification, a collection of commonly used and perceptually inspired feature extraction techniques was employed. In particular, mel-spectrogram, MFCC, and chroma-STFT representations were calculated for each audio sample to capture diverse aspects of timbre, pitch, and spectral envelope. mel-spectrograms, which map spectral energy onto the human auditory-aligned Mel-scale, have been consistently supported in literature for their effectiveness in tasks ranging from environmental sound recognition to distinguishing instruments [154]. The MFCC, obtained from the mel-spectrogram through a DCT, offers a concise description of spectral patterns and remains crucial in fields like emotion detection and respiratory sound classification [155]. Finally, the chroma-STFT features represent the distribution of spectral energy among the twelve pitch classes of Western music and serve as valuable tools for tonal and harmonic analyses.

The procedure for feature extraction utilized the open-source Python library `librosa` [156], a renowned toolbox in modern audio signal processing studies. With mel-spectrogram, MFCC, and chroma-STFT, this framework effectively captures a wide range of audio characteristics, such as spectral texture, harmonic content, and temporal evolution, thereby improving classification performance across various audio fields.

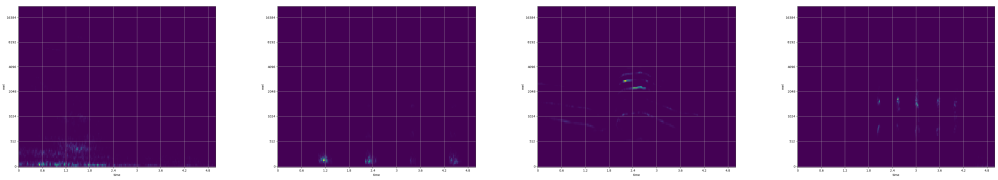
It should be emphasized that the feature extractors used in this work are based on distinct spectral representations. In particular, the mel-spectrogram and MFCC descriptors are obtained from power spectrograms, while the chroma-STFT representation is computed from the magnitude spectrum of the STFT. This methodological difference is intentional: it arises from the distinct perceptual targets and analytical roles of these features, and therefore should not be interpreted as a mistake or inconsistency in the feature design, but rather as a deliberate choice aligned with their respective objectives.

Power-oriented representations focus on how signal energy is distributed over time and frequency and on the overall dynamic range. These properties align well with human perception of loudness, which explains their widespread use in applications

such as environmental sound classification, speech processing, and AAD [41, 42, 156]. By comparison, magnitude-based chroma features are explicitly designed to downplay absolute energy levels, instead encoding the relative relationships between harmonics and pitch classes in a way that is robust to global amplitude changes [44, 157]. As a result, chroma-STFT is especially well-suited to representing tonal and harmonic content, maintaining its effectiveness even when recording conditions, playback volumes, or signal intensities vary substantially.

The decision to incorporate both power-based and magnitude-based features in this work stems from the aim of assessing deep-learning models under diverse spectral encodings that highlight different, yet complementary, acoustic properties. Comparable evaluations of heterogeneous feature sets have appeared in earlier audio classification and music information retrieval research, where the choice of representation is often guided by perceptual and task-driven considerations instead of enforcing a single, mathematically uniform description of the signal [158, 159]. In line with this perspective, the present study deliberately adopts a flexible, application-oriented feature selection approach that mirrors how features are typically chosen and combined in practical audio analysis pipelines, where robustness across varied encodings is often more important than strict theoretical consistency.

2.2.1. Mel-spectrogram



(a) airplane **(b)** breathing **(c)** cat **(d)** dog

Fig. 8. Mel-spectrogram visualization of various audio samples in the ESC-50

The utilization of mel-spectrogram becomes a foundational element for feature extraction. This methodological approach enables the conversion of raw audio signals, which exist in the time domain, into a two-dimensional representation that encompasses both time and frequency domains. This transformation allows for a more comprehensive analysis of acoustic patterns. Fundamentally, this process is anchored in the STFT, which operates by breaking down audio signals into overlapping frames before transforming these frames into the frequency domain. In particular, unlike the conventional STFT, which operates on a linear frequency scale, mel-spectrogram employs the Mel scale, a non-linear frequency representation specifically designed to correspond more closely to the perceptual characteristics of human hearing.

Initially introduced by Stevens et al. in 1937 [41], the Mel scale is the result of psychoacoustic investigations designed to translate a perceived pitch into precise

frequency values. The term 'mel' is derived from 'melody', highlighting its foundation in how we perceive sound rather than direct frequency quantification. Their research concluded that a scale compressing higher frequencies while expanding lower ones aligns more closely with human auditory perception, thereby enabling the human ear to perceive pitch variations more accurately.

Expanding on this perceptual foundation, Davis and Mermelstein (1980) [42] introduced a method to extract MFCC from the power spectrum of speech signals. Their research, centered on cepstral analysis, unintentionally established the foundation for the extensive use of Mel-filter banks in the making of mel-spectrograms. In particular, by skipping the DCT step in the extraction of MFCC, a time-frequency visualization is generated that is similar to a mel-spectrogram. This representation excels at capturing the energy distribution across frequency bands that are perceptually relevant over time, thus proving exceptionally apt for applications in speech and environmental analysis.

The process of generating a mel-spectrogram from a raw time-domain audio signal involves a sequential pipeline comprising five principal computational steps. The sequence of steps in this process, delineated according to their prevailing implementation, encompasses the application of STFT, the derivation of the power spectrogram, the implementation of the Mel-filterbank, normalization or scaling through the utilization of a predefined reference power, which culminates in logarithmic compression.

Among these, STFT serves as the foundational transformation that maps the signal into a time-frequency representation, while the power spectrogram provides a magnitude-squared view of the frequency content. The third stage, the application of the Mel filterbank, modifies the frequency axis from a linear to a perceptually motivated Mel-scale, which better aligns with human auditory perception. The subsequent step involves scaling the output by referencing a known power level. This process ensures the consistency and comparability of the amplitude values across different recordings or systems. The final step, logarithmic compression, serves to reduce the dynamic range of the spectrogram and to enhance numerical stability for further processing, such as in ML models.

However, it is noteworthy that the final two stages, namely, reference power scaling and logarithmic transformation, are occasionally regarded as optional, particularly in scenarios where a reference power value is not readily available, or where raw Mel energy values are sufficient for subsequent tasks. These variations can be attributed to the specific requirements of the application or the pre-processing protocol being employed.

A systematic mathematical formulation of each sequential computational step is presented in Equations (2.1) through (2.6). These equations detail the progression from the initial time-domain audio signal through intermediate spectral representations. Collectively, they describe the complete transformation process that yields the final

mel-spectrogram representation.

$$X[t,f] = \sum_{n=0}^{N-1} x[n]\,\omega[n-tH]\,\exp\left(-j2\pi\frac{fn}{N}\right) \tag{2.1}$$

Equation (2.1) computes the STFT of the discrete-time input signal ($x[n]$) by applying a sliding-window function ($\omega[n-tH]$) to successive, overlapping frames of the signal. This segmentation is determined by parameters such as t, frequency bin (f), discrete-time frame index (n), fast Fourier transform (FFT) size (N), and hop length (H). Each windowed segment is subsequently transformed into the frequency domain via a complex exponential kernel, yielding a time–frequency representation in the form of a spectrogram ($X[t,f]$). Each element of this complex-valued matrix encodes the spectral content corresponding to a specific f and t frame, thereby capturing the localized frequency structure of the signal over time.

$$S_{\mathrm{P}}[t,f] = |X[t,f]|^2 \tag{2.2}$$

The complex-valued output of the STFT is subsequently transformed into a power spectrogram ($S_{\mathrm{P}}[t,f]$) by computing the squared magnitude of each frequency coefficient in the resulting $X[t,f]$, as shown in Equation (2.2). This operation discards the phase information and retains only the energy content of the signal across time and frequency, yielding a real-valued matrix that characterizes the temporal evolution of spectral energy. To more closely approximate human auditory perception, the power spectrogram is then projected onto a bank of Mel-filters weight of frequency ($H_m[f]$), with the number of filters specified by Mel-filter index (m). Each triangular filter in this bank emphasizes a distinct frequency region, distributed non-linearly according to the Mel-scale. The output, denoted as the Mel-spectral function ($S_{\mathrm{mel}}[t,m]$) in Equation (2.3), represents spectral energy distributed across perceptually meaningful Mel bands, indexed by the total number of frequency bins (F), thereby compressing and smoothing the frequency resolution relative to the linear spectrum.

$$S_{\mathrm{mel}}[t,m] = \sum_{f=0}^{F-1} S_{\mathrm{P}}[t,f]\,H_m[f] \tag{2.3}$$

$$f_{\mathrm{mel}} = 2595\log_{10}\left(1+\frac{f}{700}\right) \tag{2.4}$$

As a subsequent step, Equation (2.4) defines the transformation from linear f to the Mel-scaled frequency (f_{mel}), a perceptually motivated scale that reflects the nonlinearity of human pitch perception. This logarithmic formulation provides finer frequency resolution in the lower range and coarser resolution in the higher range, thereby aligning the filterbank construction with the frequency sensitivity of the human auditory system. Optionally, the resulting $S_{\mathrm{mel}}[t,m]$ can be normalized by a reference power value, denoted as the reference power value (P_{ref}), which is

typically selected based on a predefined standard or the maximum signal power observed within the dataset. This normalization step, computed as the mel-power normalization function ($S_{\text{mel,norm}}[t,m]$) and presented in Equation (2.5), is essential for ensuring consistent amplitude scaling and for mitigating numerical instabilities in the subsequent processing stages, particularly in ML pipelines.

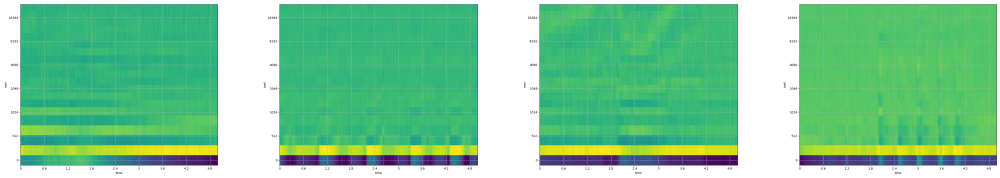
$$S_{\text{mel,norm}}[t,m] = \frac{S_{\text{mel}}[t,m]}{P_{\text{ref}}} \tag{2.5}$$

$$S_{\text{mel,log}}[t,m] = \log\left(S_{\text{mel}}[t,m] + \varepsilon\right) \tag{2.6}$$

Ultimately, a logarithmic compression is utilized on the mel-spectrogram to generate the log-mel spectrogram ($S_{\text{mel,log}}[t,m]$), as specified in Equation (2.6). This change simulates the logarithmic nature of human loudness perception, which aids in better managing the dynamic range in spectral representation. A small constant (ε) is incorporated before the logarithmic process to guarantee numerical stability and avoid undefined outcomes arising from taking the logarithm of zero.

2.2.2. MFCC

The MFCC is a widely adopted feature extraction technique in the fields of speech and audio signal processing. It transforms raw audio signals into a compact and perceptually meaningful representation that approximates the characteristics of human auditory perception. Although derived from the mel-spectrogram, the MFCC provides a more compressed representation by applying additional transformations to the Mel-frequency spectral features. At the core of this process lies the Mel-scale, a perceptually motivated frequency axis that reflects the non-linear sensitivity of the human auditory system to pitch and loudness [42, 82]. The computation of MFCC



(a) airplane **(b)** breathing **(c)** cat **(d)** dog

Fig. 9. MFCC visualization of various audio samples in the ESC-50 dataset

features typically involves five sequential steps. First, the input signal is segmented into overlapping frames and multiplied by a window function to isolate short-term time segments, thereby enabling localized frequency analysis. Second, a discrete Fourier transform is applied to each windowed frame to produce a power spectrum that characterizes the distribution of energy across frequency bins. Third, the power spectrum is passed through a bank of triangular filters spaced according to the

Mel-scale, mapping the linear frequency axis onto a perceptually relevant pitch scale. In the fourth step, the Mel-filtered spectral energies are converted to the logarithmic domain to simulate the non-linear loudness perception of the human ear. Finally, a DCT is applied to the log-Mel energies to decouple the relation of the features and compress the spectral information into a low-dimensional set of coefficients that capture the most salient perceptual characteristics. The DCT is a linear transformation that expresses a finite sequence of inputs with real values as a weighted sum of cosine basis functions oscillating at different frequencies. It is widely used in signal processing for energy compaction and feature decorrelation. The general form of the DCT for each of its coefficient (κ) is given by Equation (2.7).

$$X(\kappa) = \sum_{n=0}^{N-1} x(n) \cos\left(2\pi n \frac{\kappa}{N}\right) \tag{2.7}$$

In the context of MFCC computation, however, the DCT is not applied to the original time-domain signal. Instead, it operates on the log-Mel spectral energies obtained from Equation (2.6). This step yields the final Mel-frequency cepstral coefficients, as defined in Equation (2.8). In this formulation, the input consists of $S_{\text{mel}}[t,m]$, m, and numbers of Mel-fitlers (M), with $m \in \{0,1,\ldots,M-1\}$. The MFCC resulting coefficient (MFCC$[t,\kappa]$) provides a compact and decorrelated representation of the spectral envelope. Typically, only the lower-order coefficients are retained, as they capture the most perceptually relevant features of the signal while discarding less informative details.

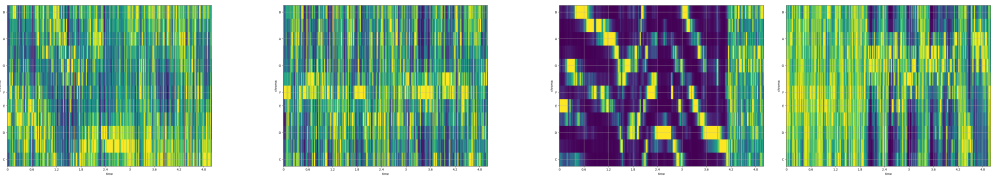
$$\text{MFCC}[t,\kappa] = \sum_{m=0}^{M-1} \log S_{\text{mel}}[t,m] \cdot \cos\left(\frac{\pi\kappa}{M}\left(m+\frac{1}{2}\right)\right) \tag{2.8}$$

Prior to this main feature extraction pipeline, raw audio signals are often subjected to pre-processing operations such as pre-emphasis filtering [160]. These procedures are designed to enhance the spectral characteristics of the signal by amplifying high-frequency components and balancing the overall energy distribution. Such pre-processing steps improve the robustness of the extracted features and contribute to enhanced performance in downstream tasks, including classification and recognition. Examples of MFCC representations obtained from pre-processed audio signals in the ESC-50 dataset are shown in Figure 9.

2.2.3. Chroma-STFT

Another widely used method for representing sinusoidal signals in the time–frequency domain, also employed in this study, is the chroma-STFT. Ellis (2007) [43] first established this approach while working with MIREX-06 beat tracking data. Later that year, Ellis and Pollner (2007) [44] refined the concept and applied it to 'cover song' detection by using 12-dimensional chroma features that capture spectral energy. Subsequently, Muller and Ewert (2011) [45] developed

software toolboxes to convert audio signals into chroma representations.



(a) airplane **(b)** breathing **(c)** cat **(d)** dog

Fig. 10. Chroma-STFT visualization of various data samples in the ESC-50 dataset

Unlike the features discussed previously, the chroma-STFT process includes segmenting a continuous audio signal into shorter segments that overlap over time. Each of these segments undergoes a FT, which facilitates a thorough spectral analysis over time. In this context, 'chroma' refers to the twelve semitone pitch classes, such as C, C#, D, etc., that exist within a single octave, particularly in Western musical traditions, and are perceptually significant in numerous musical scenarios. The chroma-STFT approach aligns the frequency components of the signal with these pitch classes to encapsulate harmonic and melodic information while disregarding absolute pitch levels [83].

This chroma-focused mapping has shown advantages in areas such as music information retrieval, including chord recognition, key detection, structural segmentation, and identifying cover versions of songs [161]. By directing spectral energy into chroma categories, the technique emphasizes harmonically significant features, providing a simplification of a spectral content that maintains perceptual relevance [162, 163]. Moreover, it allows for the tracking of harmonic changes over time, which is beneficial for analyzing signals that lack stationarity or display dynamic variations [164]. In contrast to full-spectrum representations, chroma-STFT preserves tonal structures more in sync with human auditory comprehension, thus enhancing the interpretability and effectiveness of subsequent audio analysis processes [83, 163].

To implement chroma-STFT in practice, a sequence of signal processing operations is used. The process begins with the application of the STFT, which is performed by using the `Librosa` library in Python. The input signal is segmented into overlapping frames, and a Hanning window is applied to each segment. The resulting $X[t,f]$ is then converted into a magnitude spectrogram ($S_{\text{mag}}[t,f]$). Since $X[t,f]$ has a complex characteristic, its magnitude is computed by performing the Euclidean norm of the real and imaginary components, as shown in Equation (2.9).

$$S_{\text{mag}}[t,f] = |X[t,f]| = \sqrt{\text{Re}\,[X[t,f]]^2 + \text{Im}\,[X[t,f]]^2} \tag{2.9}$$

Following this, the linear frequency bins are mapped to pitch classes using a logarithmic transformation. This mapping procedure is performed by utilizing Equation (2.10), where each f is converted relative to a fixed reference frequency

(f_{ref}), which is typically set at 440 hertz (Hz) and is denoted as A4. The mod 12 operation ensures octave equivalence, assigning all harmonically related frequencies to the same chromatic class.

$$p(f) = \left[12 \times \log_2 \left(\frac{f}{f_{\text{ref}}} \right) \right] \mod 12 \tag{2.10}$$

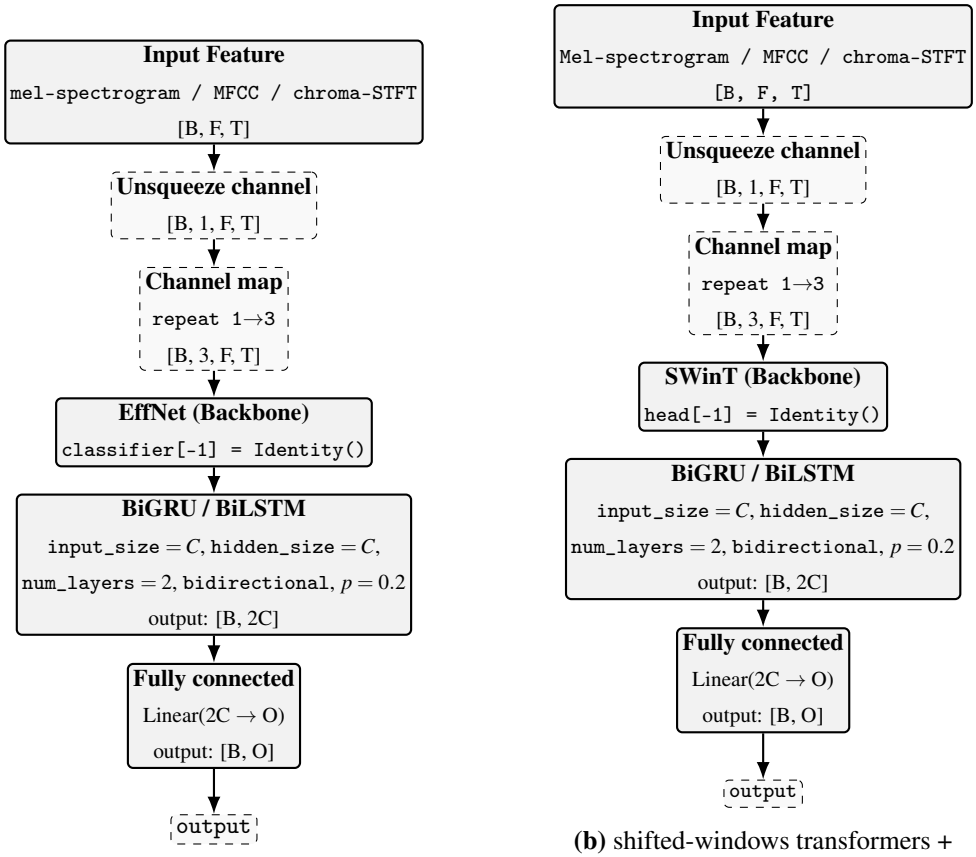
$$S_{\text{chroma}}[t,p] = \sum_{f \in p} S_{\text{mag}}[t,f] \tag{2.11}$$

The final step aggregates the spectral energy into chroma bins. For each t and chroma bin (p), chroma energy ($S_{\text{chroma}}[t,p]$) is calculated by adding the magnitude values of all frequencies that correspond to the same chroma bin, as shown in Equation (2.11). Figure 10 presents a visual comparison of chroma-STFT representations for various audio recordings drawn from distinct object categories. Unlike conventional spectrograms, which depict energy over linear frequency bands, the chroma representation organizes information by the pitch class. This structural distinction enables the chroma-STFT to highlight harmonic structures and pitch constellations that are often obscured in linear frequency-based analyses.

Although mel-spectrograms and MFCC features are commonly utilized for tasks such as instrument and speaker identification, chroma-STFT is particularly well-suited to capture harmonic and tonal characteristics. However, its utility extends beyond musical applications. In fact, it has shown relevance in broader audio classification domains. As with other feature representations used in this study, chroma-STFT is derived from the framework STFT. However, it is distinguished by its grounding in music-theoretic principles and its ability to reveal harmonic similarities across temporally varying audio segments.

2.3. Proposed Models

This study introduces several hybrid DL frameworks designed to better capture temporal patterns in acoustic signals by combining RNN modules with pre-trained models. We investigate four architectures (i.e., EffNet-BiGRU, EffNet-BiLSTM, SWinT-BiGRU, SWinT-BiLSTM) that pair a CNN-based EffNet or attention-driven SWinT feature extractor with a bidirectional recurrent layer (BiGRU or BiLSTM) at the final classification stage (see Fig. 11). The design is modular: the backbone's classifier is removed, preserving its spatial and hierarchical representation learning, while the recurrent block adds sequential memory, which is important for decisions where fine spectral variations over short intervals are semantically relevant. Fusing spatial and temporal features only in the last layers is particularly effective under noisy, overlapping acoustic conditions, such as industrial monitoring and environmental audio. The framework generalizes to diverse pretrained backbones and datasets, enabling domain-specific fine-tuning. We evaluate these hybrids on multiple benchmarks, including industrial anomaly detection (MIMII), environmental sound classes (ESC-50), and natural acoustic scenes (FSC22). Prior work [165–167] has



(a) EfficientNet + BiGRU (EffNet-BiGRU) / EffNet-BiLSTM

(b) shifted-windows transformers + BiGRU (SWinT-BiGRU) / SWinT-BiLSTM

Fig. 11. Proposed models

shown that the addition of RNN layers enhances sequence modeling and improves robustness and accuracy in complex auditory tasks.

The proposed hybrid architectures, illustrated in Figure 11, operate on time–frequency representations (i.e., mel-spectrogram, MFCC, and chroma-STFT) with shape $[B, F, T]$. We first introduce a channel dimension to obtain $[B, 1, F, T]$, then replicate it to $[B, 3, F, T]$ so that the inputs are compatible with ImageNet-pretrained models. Two backbones are employed: (1) a CNN-based EffNet and (2) SWinT with custom self-attention windows. Their feature extractors are kept intact, but the final classification layers are replaced with identity mappings (`classifier`$[-1]$ = `Identity()` for EffNet and `head`$[-1]$ = `Identity()` for SWinT). Each backbone outputs a feature vector of size `C`, which is then fed into a bidirectional recurrent head implemented with either BiGRU or BiLSTM, configured with `input_size` $= C$, `hidden_size` $= C$, `num_layers` $= 2$, `bidirectional = True`, and $p = 0.2$. This

recurrent head produces 2C features, which are mapped by a fully connected layer $\texttt{Linear}(2C \rightarrow K)$ to class logits $[B, O]$. During training, cross-entropy is applied directly to the logits (without softmax), while softmax is only used at inference to obtain probabilities. This design yields four model variants (i.e., EffNet-BiGRU, EffNet-BiLSTM, SWinT-BiGRU, and SWinT-BiLSTM), which share the same recurrent classification head but differ in their backbone, thus combining sophisticated spatial/hierarchical feature extraction with temporal modeling.

The evolution of audio processing strategies has been strongly shaped by foundational work underscoring the effectiveness of RNN-based layers. Etienne et al. (2018) [168] introduced a CNN–LSTM architecture for SER, achieving weighted and unweighted accuracies of 64.5% and 61.7%, respectively. Later, Zhang and Martinez-Garcia (2020) [169] employed an LSTM network to identify bearing operating conditions from emitted audio, reporting a 94.7% classification accuracy. In 2025, Inapagolla and Babu [170] presented a hybrid system that fuses CNN, RNN, and LSTM models for speaker recognition in audio fingerprinting, reaching 90% accuracy while improving computational efficiency. That same year, Qurthobi et al. [148] proposed a CNN-BiLSTM model to classify audio transcripts from the FSC22 dataset. Together, these studies demonstrate the strong promise of RNN layers for recognizing and classifying sequential patterns across diverse audio processing tasks.

2.4. *k*-folds CV

The *k*-folds CV is a type of cross-validation (CV) that is widely used and observationally tested model evaluation method intended to assess the applicability of ML algorithms with statistical reliability. This technique proves especially advantageous when dealing with a limited-size dataset, offering a more robust and unbiased assessment than a simple train-test split [171]. Using of each data point for both training and validation reduces sampling bias and reduces the risk of overfitting to particular data subsets. Furthermore, this approach enables the validation of the observations by altering the sets of data used for training and validation.

In the implementation of *k*-folds CV, the original dataset ($\texttt{D}$) is divided into arbitrary integer number (k) segments, known as $\texttt{D}_k$s, that are roughly equal in size and do not overlap. Equation (2.12) describes this partitioning process mathematically, ensuring that all subsets collectively reassemble the original dataset. In each round of the *k*-folds CV procedure, a different $\texttt{D}_k$ is designated as the validation set, while the remaining segments are merged to form the training set. This process is repeated according to the total folds (K), allowing each fold to be used once as a validation set and then become part of the training set in the subsequent rounds. As a result, the model undergoes K cycles of training and validation, each employing different dataset partitions. This systematic rotation of folds enables a thorough performance evaluation by accounting for the variation within different subsets during training and testing. Furthermore, every single performance metric on the k-th validation fold ($\mathscr{M}_k$) (e.g., accuracy, loss, and recall, etc.) is computed and collected for each validation

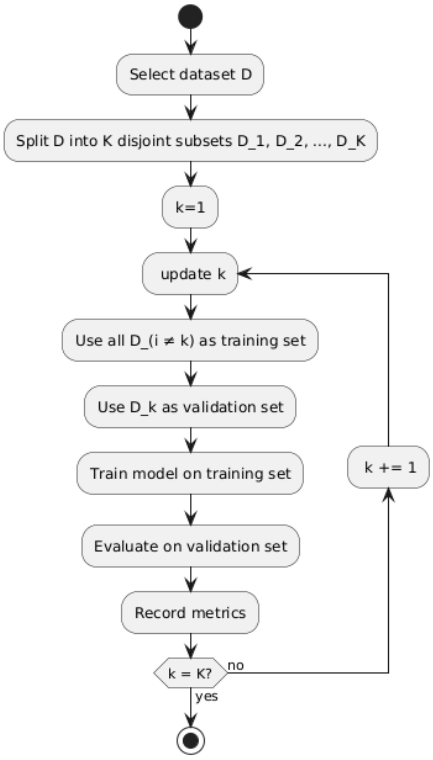


Fig. 12. Steps for implementing k-fold cross-validation

set. For visualization purposes, all average performance metrics ($\bar{\mathscr{M}}$) and deviation standards of each performance metric ($\sigma_{\mathscr{M}}$) are then calculated using Equations (2.13) and (2.14), respectively. Figure 12 visually represents the entire process as applied in k-folds CV.

$$\mathtt{D} = \bigcup_{k=1}^{K} \mathtt{D}_k, \quad \mathtt{D}_i \cap \mathtt{D}_j = \emptyset \text{ for } i \neq j \tag{2.12}$$

$$\bar{\mathscr{M}} = \frac{1}{K} \sum_{k=1}^{K} \mathscr{M}_k \tag{2.13}$$

$$\sigma_{\mathscr{M}} = \sqrt{\frac{1}{K-1} \sum_{k=1}^{K} \left(\mathscr{M}_k - \bar{\mathscr{M}}\right)^2} \tag{2.14}$$

To ensure that the process proceeds effectively, selecting an appropriate value of K proves beneficial. Consequently, numerous datasets often provide recommendations on the optimal number of K to employ. Piczak, as the sole creator of ESC-50, has endorsed $k = 5$ since its inception for consistent evaluations [5, 172]. As a result,

Zhou et al. (2025) [172] adopted a similar k value to test a new CNN-based model on ESC-50, achieving competitive performance across multiple folds. On the other hand, Zhang et al. (2018) [173] suggested $k = 10$ during the inspection of the UrbanSound8k [6] dataset to produce better classification results. This assertion was further substantiated by Demir et al. (2020) [174], who reported accuracy means of 84.20% and 86.70% with $k = 5$ and $k = 10$, respectively. Meanwhile, the FSC22 dataset by Bandara et al. (2023) [7] has provided no specific information regarding the best value of k. Nonetheless, research by Ranmal et al. (2024) [143] promoted the implementation of $k = 5$ for optimal results.

In this study, we applied a uniform five-fold CV strategy to all three datasets (i.e., MIMII, ESC-50, and FSC22), using the folds exclusively for training and validation and not defining a separate test set. This decision was driven by the limited number of samples, particularly in ESC-50, whose original five-fold split design provides only eight samples per class for validation or testing [5]. A unified five-fold CV scheme ensured methodological consistency and maximized use of the scarce data. Prior work has shown that, in low-data scenarios, evaluating models without a distinct test set and relying instead on CV for both training and performance estimation can still yield dependable generalization estimates [175]. Similarly, nested CV is recommended when data scarcity precludes separate validation and test sets, as it supports effective hyperparameter optimization and unbiased evaluation when using only the available data.

A central difficulty in applying k-folds CV is selecting the most suitable variant for training and validation. Among the available choices, stratified k-folds CV is most commonly adopted because each fold preserves the overall class distribution, which is especially important for imbalanced datasets. As noted by Hastie et al. (2009) [176], this stratification enhances the reliability of evaluation metrics, particularly in audio classification tasks where certain sound classes may be underrepresented. More broadly, k-folds CV is fundamental to audio ML frameworks: it maximizes the utility of limited data, supports robust training and validation, and enables fair model comparison, hyperparameter tuning, and generalization error assessment. Combined with stratification and separate holdout test sets, it provides a solid basis for performance evaluation in environmental, industrial, and ecological audio applications.

2.5. Workflow

Figure 13 summarizes the overall workflow used in this study. At its core is the deliberate choice and combination of key methodological components (i.e., datasets, feature extraction techniques, and DL frameworks) which together aim to deliver strong classification performance while maintaining robustness to diverse acoustic conditions. The workflow begins with the construction of curated datasets, continues with the extraction of relevant acoustic features, and ends with their processing by selected DL models, with each stage designed to preserve data quality and ensure computational efficiency. To further enhance the reliability of the results, the study

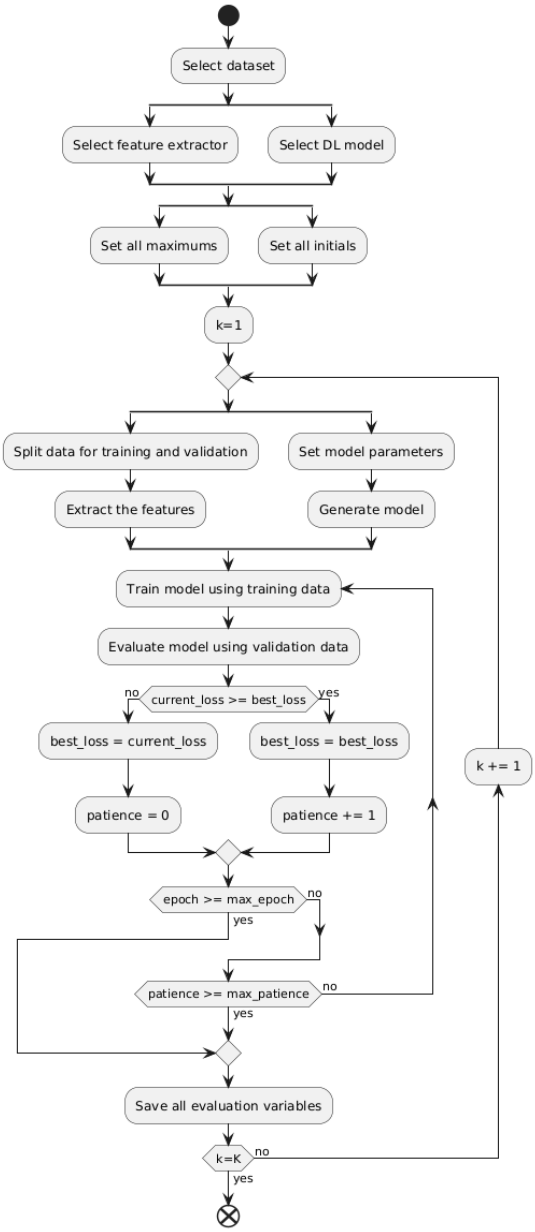


Fig. 13. Workflow procedures

uses k-folds CV to support model generalization and early stopping based on validation loss to reduce overfitting. These elements jointly establish a rigorous, reproducible experimental setup that yields trustworthy results across multiple evaluation scenarios.

2.5.1. Dataset selection

As previously discussed, the selection of datasets is crucial for assessing the effectiveness of audio classification systems. In this study, we utilize three unique and commonly used environmental sound datasets: ESC-50, MIMII, and FSC22. Each dataset is specifically chosen to represent particular real-world acoustic settings, with ESC-50 focusing on residential sounds, MIMII capturing industrial mechanical noises, and FSC22 illustrating natural forest environments. This variety in environmental contexts enables a comprehensive evaluation of model performance across diverse acoustic scenarios, thereby boosting the ecological relevance of the results. Apart from their environmental focus, these datasets vary significantly in structural features, such as the number of audio samples, class definitions, and recording environments. The MIMII dataset, in particular, presents a significant class imbalance, with certain machine types and operational scenarios overrepresented. This imbalance poses challenges in training DL models, as it can bias the learning process and affect model generalization. The resolution of these issues is vital to reaching firm and unbiased conclusions, making it a key part of the methodological strategy of the study.

2.5.2. Feature extraction techniques

The selection of appropriate acoustic feature representations is also vital in determining the efficacy, clarity, and computational feasibility of audio classification systems. This research employs three widely acknowledged feature extraction techniques: mel-spectrogram, MFCC, and chroma-STFT. These features are extensively cited in the literature for their ability to capture perceptually significant audio signal representations. The goal of incorporating different feature types is to evaluate their unique impacts on classification accuracy, training velocity, and computational requirements. Specifically, the mel-spectrogram offers comprehensive time-frequency representations, MFCC delivers succinct spectral summaries that align with human auditory perception, and chroma-STFT focuses on pitch class details beneficial in tonal contexts. These representations vary in their dimensional complexity and informational richness, which influence the model performance as well as the time and memory necessary for training. While high-resolution features may provide more information, they demand substantial computational resources, posing challenges in real-time or resource-constrained settings. By meticulously examining these trade-offs, the study contributes to a better understanding of how feature selection impacts the efficiency and precision of DL-based audio classification systems.

In exploring the implementation of feature extraction techniques, it is crucial to scrutinize their default operational parameters, as outlined by the well-known `librosa` library. This understanding provides insights into the fundamental similarities across functions, while highlighting the structural assumptions each feature type embodies. Illustrated in Table 11, several significant default configurations are common among the three extractors in the Librosa library, such as a sampling rate (`sr`) of 22050 Hz,

Table 11. Comparison of default parameters in mel-spectrogram, MFCC, and chroma-STFT with Librosa library

Parameter	`melspectrogram()`	`mfcc()`	`chroma_stft()`
`sr`	22050	22050	22050
`n_fft`	2048	2048	2048
`hop_length`	512	512	512
`win_length`	None	None	None
`window`	'hann'	'hann'	'hann'
`center`	True	True	True
`pad_mode`	'reflect'	'reflect'	'reflect'
`power`	2.0	– (calls internally)	–
`n_mels`	128	128 (used inside)	–
`fmin`	0.0	0.0 (used inside)	0.0
`fmax`	`sr`/2	`sr`/2 (used inside)	`sr`/2
`n_mfcc`	–	20	–
`n_chroma`	–	–	12
`dct_type`	–	2	–
`norm`	None	'ortho'	None
`htk`	False	False (used inside)	False
`ref`	–	1.0 (used in log-mel scaling)	–

an FFT windows size (`n_fft`) of 2048, a hop length (`hop_length`) of 512, and the use of the Hann window function to smoothen the spectral analysis. These shared defaults ensure consistent feature extraction, aiding in experimental setups and model refinement for comparative studies. Despite these apparent similarities, each method employs unique transformations and spectral focuses to capture specific perceptual or musical features. Moreover, the output sizes of each function's second dimension differ. For example, assuming a mono-audio signal with identical input across functions, the extraction outputs will vary. In a mel-spectrogram, the entries in the second dimension correspond to the number of mel bands (`n_mels`), while in MFCC and chroma-STFT, they match the number of MFCCs to return (`n_mfcc`) and the number of chroma bins to produce (`n_chroma`), respectively. Thus, the mel-spectrogram provides more granularity than others, but is computationally burdensome and requires greater resources. Understanding these configurations is essential for ensuring experiment reproducibility and clarity, especially since minor preprocessing changes can significantly impact classification outcomes. This concise table serves as a crucial resource for understanding the parameters within which these audio representations function.

2.5.3. Cross-validation strategy and early stopping

The selection of a suitable architecture DL and feature extraction techniques is crucial, but the implementation of *k*-folds CV and the early stop mechanism is also a beneficial methodological aspect in the design and evaluation of audio classification

systems. k-folds CV's main role is to boost the reliability and reproducibility of experimental results by presenting potential biases from arbitrary splits in training and validation datasets. This technique ensures that every data sample participates equally in both training and evaluation, thereby promoting consistency and fairness in performance measurement across different folds. Studies by Piczak (2015) [5] and Ranmal et al. (2024) [143] have advocated for using $k = 5$ in ESC-50 and FSC22, respectively, due to its balance between computational efficiency and statistical soundness. In line with these recommendations and with the aim of methodological coherence, this study used a 5-fold CV approach for all experiments. Despite the absence of predefined folds in the MIMII dataset, applying k-folds CV here establishes a consistent comparative framework with ESC-50 and FSC22, allowing for more generalizable conclusions about the model performance across diverse acoustic environments.

2.5.4. Training parameters

Table 12. The values of all parameters and limits during training and validation

Parameter	Value	Note
Sampling rate	`None`	Adjust to the default value of the dataset
Maximum fold	5	Number of training and validation cycles
Maximum epoch	2000	Maximum epoch value for training the model
Maximum patience	10	Number of waits before stopping due to no improvement
Learning rate	10^{-6}	Controls model's weight update size
Batch size	32	Samples processed before weight update
Loss criteria	Cross Entropy	Function to measure prediction error during model training
Optimizer	AdamW	Weight adjuster to minimize loss function

To maintain a consistent and reliable training setup, the final stage of the experimental design defines key hyperparameters and optimization strategies that govern model convergence and generalization. These settings are essential for stabilizing training behavior and ensuring comparable evaluations across different experimental configurations. Although the choice of the dataset, feature extraction method, and model architecture forms the core of the audio classification pipeline, the training setup introduces pivotal controls that shape learning efficiency and final accuracy. These include the number of training folds, the maximum number of training epochs, and the early stopping patience, which together help prevent overfitting and support stable convergence. Additionally, decisions regarding the learning rate, batch size, optimizer, and loss function strongly influence gradient updates, by trading off speed against stability. In this work, adaptive optimizers such as AdamW, combined with standard loss functions like cross-entropy, provide robust performance across varying feature sets and datasets. Parameter values and

defaults are carefully tuned using preliminary experiments; they follow the current best practices reported in the literature. For transparency and reproducibility, Table 12 summarizes the parameter ranges and final values used for training and validation in this study.

2.5.5. Evaluation metrics

As previously described in Chapter , this study adopts a set of complementary evaluation metrics to provide a thorough and reliable assessment of the predicted classifications. In particular, five standard measures (i.e., accuracy, AUC, precision, recall, and F1) are employed to systematically quantify the performance of the classification models from different perspectives [177, 178]. Together, these metrics capture both overall correctness and the balance between correctly and incorrectly identified instances.

$$\text{Accuracy} = \frac{\mathscr{T}_+ + \mathscr{T}_-}{\mathscr{T}_+ + \mathscr{T}_- + \mathscr{F}_+ + \mathscr{F}_-} \tag{2.15}$$

$$\text{AUC} = \int_0^1 \mathscr{T}_{+R}(\mathscr{F}_{+R}) d(\mathscr{F}_{+R}) \tag{2.16}$$

$$\text{Precision} = \frac{\mathscr{T}_+}{\mathscr{T}_+ + \mathscr{F}_+} \tag{2.17}$$

$$\text{Recall} = \frac{\mathscr{T}_+}{\mathscr{T}_+ + \mathscr{F}_-} \tag{2.18}$$

$$\text{F1} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\mathscr{T}_+}{2\mathscr{T}_+ + \mathscr{F}_+ + \mathscr{F}_-} \tag{2.19}$$

The mathematical definitions of these measures are given in Equations (2.15) through (2.19), which specify in sequence how each metric is computed. These equations explicitly show how the values of accuracy, AUC, precision, recall, and F1 are derived from the fundamental components of the confusion matrix, namely, true positive ($\mathscr{T}_+$), true negative ($\mathscr{T}_-$), false positive ($\mathscr{F}_+$), and false negative ($\mathscr{F}_-$). This formalization ensures that the evaluation procedure is transparent, reproducible, and consistent across all models examined in this study.

$$\mathscr{T}_{+R} = \frac{\mathscr{T}_+}{\mathscr{T}_+ + \mathscr{F}_-} \tag{2.20}$$

$$\mathscr{F}_{+R} = \frac{\mathscr{F}_+}{\mathscr{T}_- + \mathscr{F}_+} \tag{2.21}$$

In addition, the calculation of AUC in Equation 2.16 depends on the true positive rate ($\mathscr{T}_{+R}$), and false positive rate ($\mathscr{F}_{+R}$), which are formally defined in Equations (2.20) and (2.21), respectively. By explicitly relating these quantities to the individual cells of the confusion matrix, the metric definitions make transparent how AUC condenses a classifier's behavior over a continuum of operating points. This explicit

linkage emphasizes that AUC reflects performance across varying decision thresholds, rather than characterizing the classifier at only a single, fixed cutoff. As a result, these formulations provide a more comprehensive and fine-grained characterization of model performance, facilitating a deeper understanding of how changes in the decision threshold influence the trade-offs between true and false classifications [179, 180].

2.6. Working Environment

2.6.1. Hardware

The experiments presented in this study utilized the hardware setup detailed in Table 13, maintaining a consistent and reproducible computational setting. This system was driven by a high-performance Intel Core i9-14900K processor along with 128 GB of RAM, and was further enhanced by the NVIDIA GeForce RTX 5090 GPU for accelerating deep learning tasks. The configuration included two NVMe solid-state drives, Samsung 990 PRO and WD Black SN770, each providing a 2 TBs capacity to enable high-speed data transfer and effective storage management, crucial for managing extensive audio datasets. Networking reliability for dataset acquisition and software updates was achieved through a Realtek 2.5 GbE Ethernet controller, and Intel Wi-Fi, while audio playback was supported by onboard and GPU-integrated audio controllers for accurate signal verification.

Table 13. Summary of system hardware specifications

Component	**Specification**
Processor (CPU)	Intel Core i9-14900K, 24 cores (Hyperthreading), up to 5.7 GHz Turbo
Memory (RAM)	128 GB DDR4/DDR5 (system memory)
Graphics (GPU)	NVIDIA GeForce RTX 5090 (GB202, driver: Nvidia)
Storage	Samsung 990 PRO 2TB NVMe SSD WD Black SN770 2TB NVMe SSD
Networking	Realtek RTL8125 2.5 GbE Ethernet (active, 1 Gbps link) Intel Raptor Lake CNVi Wi-Fi (inactive)
Audio	NVIDIA HDMI/DP Audio (GPU integrated) Intel Raptor Lake HD Audio Controller (onboard)

2.6.2. Software

The computational environment, summarized in Table 14, was set up on `CentOS Stream 10`, an OS from the `Red Hat Linux` family that emphasizes long-term stability and consistent maintenance, making it well-suited to demanding scientific workloads. Within this OS, the NVIDIA CUDA framework enabled GPU-accelerated parallelism, substantially speeding up model training and supporting large-scale numerical experiments. This combination of a stable, production-ready software stack with high-performance hardware acceleration provided a solid basis for all

Table 14. Software platform used for implementing and evaluating the DL models

Category	Software / Library	Purpose
Operating system (OS)	CentOS 10 Stream Red Hat Linux Family	Execution environment for model development, training, and evaluation
Programming language	Python (v3.x)	Primary language for data processing, feature extraction, and model implementation
DL framework	PyTorch	Design and training of models
Audio feature extraction	Librosa	Computation of mel-spectrogram, MFCC, and chroma-STFT representations
Numerical computation	NumPy, SciPy	Efficient numerical operations and signal-processing utilities
ML utilities	scikit-learn	Performance metrics, CV procedures, and statistical evaluation
Visualization	Matplotlib, Seaborn	Visualization of waveforms, spectrograms, confusion matrices, and training dynamics
Hardware acceleration	CUDA, cuDNN	GPU-accelerated training and inference of DL models

experimental phases, from early prototypes to large-batch runs and performance assessments.

A dedicated and well-integrated software stack was configured to guarantee consistency, adaptability, and high computational performance across every stage of the experimental pipeline. Python 3 served as the central development platform, offering a broad ecosystem of scientific computing and data analysis packages that enabled the design of reliable data pre-processing pipelines, sophisticated feature engineering procedures, and systematic data validation steps. These libraries made it possible to implement both standard and custom workflows while maintaining clear, maintainable code structures.

For the development of DL architectures, the `PyTorch` framework was selected, primarily due to its dynamic computation graph paradigm, intuitive and expressive API, and extensive support for GPU-accelerated operations. These characteristics are particularly advantageous for iterative experimentation, as they simplify the process of modifying model components, integrating novel layers or loss functions, and monitoring intermediate outputs during training. Furthermore, `PyTorch` integrates well with complementary tools for visualization, automatic differentiation, and deployment, which further streamlined the research process.

When used together, Python 3 and `PyTorch` enabled a highly modular, extensible approach to model design, allowing rapid prototyping of alternative network topologies and systematic comparison of competing architectures under controlled conditions. This combination also facilitated efficient debugging, fine-grained performance profiling, and the incorporation of automated testing

routines. In parallel, the environment was tightly coupled with version control systems and experiment tracking utilities, so that configuration files, model checkpoints, and evaluation metrics could be recorded and revisited in a transparent manner. Overall, this software ecosystem ensured that all neural network training and evaluation procedures in this study were performed in a reproducible, reliable, and computationally optimized setting, providing a solid foundation for rigorous experimental analysis.

3. PERFORMANCE EVALUATION AND ANALYSIS

This chapter reports an experimental assessment of the proposed audio classification frameworks across varied acoustic settings, spanning industrial, urban, and natural soundscapes. Several public datasets are used to evaluate model robustness and generalization under heterogeneous, noisy conditions. Although anomaly event detection remains the core objective of this dissertation, benchmark datasets such as ESC-50 and FSC22, which are typically framed as multi-class classification problems, are instead used as proxy testbeds to investigate anomaly-related characteristics, including resilience to background noise, detection of rare or short-duration events, and consistency across diverse sound classes.

Model performance is evaluated with standard classification metrics—accuracy, precision, recall, F1-score, and AUC. Accuracy is included mainly for completeness and comparison with prior studies (see Section 3.1), while interpretation focuses on anomaly-centered metrics that better capture detection robustness under class imbalance and noise. Loss values are reported as-is and used only to track training dynamics and determine early stopping. All experiments use a unified protocol that combines early stopping with k-folds CV ($k = 5$) across datasets and model variants to support statistically reliable results.

The presentation of the experimental results is organized as follows. Section 3.1 describes the primary classification findings, including accuracy trends and training behavior. Section 3.2 provides a detailed analysis of model performance according to different evaluation metrics. Finally, Section 3.3 synthesizes broader insights on metric-specific patterns, comparisons with previous work, and other central observations.

3.1. Classification Results

3.1.1. EffNet

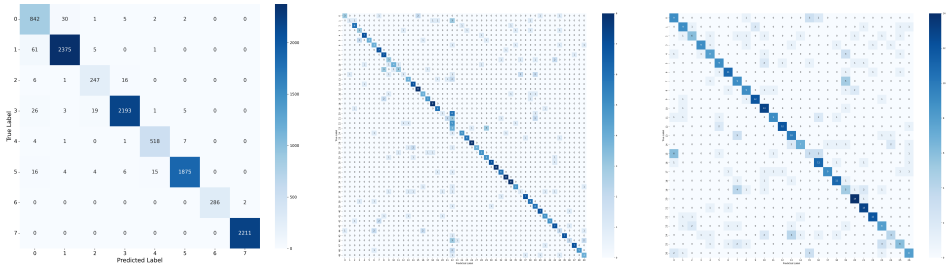
The classification outcomes of the EffNet model, assessed using the MIMII dataset and triple audio feature representations (i.e., chroma-STFT, MFCC, and mel-spectrogram) are comprehensively summarized in Table 15. This evaluation spans five k-folds CV iterations. Detailed performance metrics are highlighted, focusing on accuracy expressed as percentages and loss in the raw form. The last two rows of the table provide a consolidated view of these metrics through $\bar{\mathscr{M}}$ and $\sigma_{\mathscr{M}}$ calculated across all folds. Furthermore, Figure 14 unveils the confusion matrices of the highest classification achievements for every dataset using EffNet.

Table 15 displays that the mel-spectrogram stands out by achieving the highest accuracy in each fold, with figures spanning between 96.89% and 97.74%, culminating in an impressive average accuracy of 97.43% and a minimal standard deviation of 0.35%, thereby underscoring both excellent efficiency and consistency across folds. The peak accuracy classification results are illustrated in Figure 14a via the confusion matrix of EffNet. Nevertheless, an irregularity was detected in the fifth fold, where loss values surged to 15.29, thus indicating that the combination of

Table 15. Accuracy (in %) and loss scores for MIMII dataset's classification using EffNet

k	Chroma-STFT		MFCC		Mel-spectrogram	
	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss
1	58.71	1.34	89.19	0.38	97.60	0.10
2	57.87	7.50	89.09	0.35	97.26	0.12
3	47.49	16.13	89.41	0.38	97.66	0.12
4	58.71	40.44	89.78	0.37	97.74	0.10
5	59.51	4.65	90.21	0.38	96.89	15.29
$\bar{\mathscr{M}}$	56.46	14.01	89.54	0.37	97.43	3.15
$\sigma_{\mathscr{M}}$	5.05	15.76	0.46	0.01	0.35	6.79

training and validation data in this fold produced an undesired situation where the early stop mechanism failed to prevent overfitting. Such isolated loss anomalies are attributed to unfavorable validation splits rather than systematic optimization failure, and do not materially affect cross-validated performance trends. Conversely, the chroma-STFT representation exhibited inferior performance, achieving the lowest average accuracy of 56.46% with a large variance of 5.05%, signaling unreliability within this anomaly detection context. Moreover, chroma-STFT portrayed highly erratic loss statistics, with an average of 14.01 coupled with a substantial standard deviation of 15.76, further accentuating its inconsistency. In comparison, MFCC features demonstrated stable and robust results, exhibiting consistent accuracy with $\bar{\mathscr{M}} = 89.54\%$ and $\sigma_{\mathscr{M}} = 0.46\%$, and maintaining comparatively stable loss figures with $\bar{\mathscr{M}} = 0.37$ and $\sigma_{\mathscr{M}} = 0.01$, albeit slightly elevated relative to the mel-spectrogram.



(a) MIMII **(b)** ESC-50 **(c)** FSC22

Fig. 14. Confusion matrix during the highest achievement of classification with EffNet

Table 16 elaborates on the performance metrics for the EffNet model, as it was tested on the ESC-50 dataset across five distinct k-folds CV folds with three unique types of audio features. Displayed metrics include both accuracy and loss, expressed in percentage and raw terms, respectively. The final two rows summarize the data,

Table 16. Accuracy (in %) and loss scores for ESC-50 dataset's classification using EffNet

k	Chroma-STFT		MFCC		Mel-spectrogram	
	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss
1	27.50	2.58	56.75	1.53	60.50	1.47
2	30.00	2.56	46.25	2.12	56.50	1.50
3	25.75	2.66	56.25	1.65	56.50	1.60
4	29.00	2.53	59.25	1.41	59.25	1.50
5	26.00	2.74	56.75	1.60	62.25	1.40
$\bar{\mathscr{M}}$	27.65	2.61	55.05	1.66	59.00	1.49
$\sigma_{\mathscr{M}}$	1.85	0.09	5.06	0.27	2.52	0.07

presenting the average metrics ($\bar{\mathscr{M}}$) and their standard deviation ($\sigma_{\mathscr{M}}$). From the findings, it is apparent that mel-spectrogram features consistently deliver superior results, achieving the highest accuracy in each fold, ranging between 56.50% and 62.25%, with an overall mean accuracy of 59.00% and a standard deviation of 2.52%, reflecting elevated performance reliability across all folds. The peak performance of the model is represented in Figure 14b through a confusion matrix. In terms of loss, EffNet, when paired with mel-spectrogram features, demonstrates less fluctuation compared to other features. Conversely, the chroma-STFT features register the least accurate classification results, with an average accuracy of 27.65%, highlighting its inconsistent performance and lack of effectiveness in identifying industrial anomalies in this instance. Despite exceeding chroma-STFT in accuracy with a mean of 55.05% and a standard deviation of 5.06%, MFCC features reveal significant variability in loss metrics. Although showing a lower mean loss score relative to chroma-STFT, MFCC endures the highest standard deviation in loss, indicating a low degree of consistency.

Table 17. Accuracy (in %) and loss scores for FSC22 dataset's classification using EffNet

k	Chroma-STFT		MFCC		Mel-spectrogram	
	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss
1	35.06	2.31	58.27	1.37	45.93	1.80
2	31.85	2.79	60.25	1.38	49.63	1.66
3	28.64	2.55	60.25	1.37	52.35	1.57
4	32.84	2.34	56.54	1.51	46.67	2.12
5	27.41	2.48	34.57	3.08	48.40	1.70
$\bar{\mathscr{M}}$	31.16	2.49	53.98	1.74	48.60	1.77
$\sigma_{\mathscr{M}}$	3.12	0.19	10.96	0.75	2.55	0.21

The classification performance of the EffNet model on the FSC22 dataset, as presented in Table 17, reveals relatively modest results across all feature extraction methods, reflecting the inherent difficulty of classifying acoustically diverse and

unstructured sounds typical of wilderness environments. Unlike MIMII and ESC-50, MFCC achieves the highest mean accuracy at 53.98% with EffNet, though with a notably high standard deviation of 10.96%, indicating significant variability across folds and potential instability in generalization.

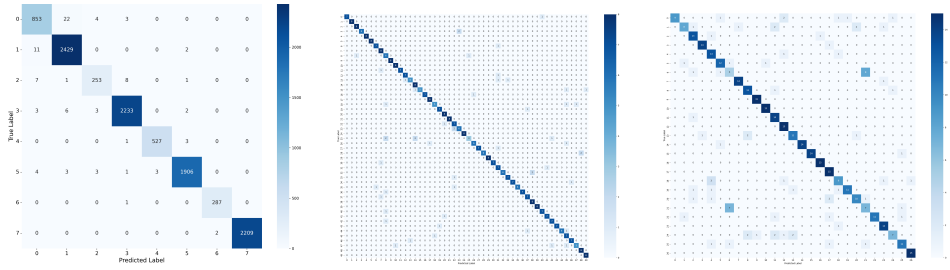
Figure 14c reveals the confusion matrix during its highest achievement. Unfortunately, although the combination of MFCC and EffNet yields the lowest average loss of 1.74, it is accompanied by a substantial deviation of 0.75, further suggesting susceptibility to inconsistent convergence. The main cause of this situation is evidenced in Table 17 by looking at the products of the combination of MFCC and EffNet during the fifth cycle of the implementation of k-folds CV, where 34.57% and 3.08 represent its accuracy and loss scores, respectively. In contrast, the mel-spectrogram features yield a slightly lower average accuracy of 48.60%, but exhibit much greater stability with a standard deviation of only 2.55%. Although they produce relatively higher average losses, the mel-spectrogram features provide a lesser variance than MFCC, suggesting that although they may not produce the highest accuracy, they offer more reliable convergence behavior. Chroma-STFT performs the worst overall in the evaluated non-tonal settings, with the lowest average accuracy (31.16%) and the highest average losses profile with average score of 2.49 and standard deviation of 0.19. These findings emphasize the challenge of applying baseline convolutional models such as EfficientNet to datasets such as FSC22, where environmental sounds are less structured, highly variable, and often noisy. Although MFCC delivers the highest accuracy, the lower variance of the results based on the mel-spectrogram may make it a more practical choice for stable application, particularly in tasks requiring robustness over optimality.

3.1.2. SWinT

Table 18 offers a detailed overview of the classification outcomes of the SWinT model evaluated on the MIMII dataset, utilizing five k-folds CV folds and three distinct audio feature types. In comparing MIMII dataset results between SWinT and EffNet, the SWinT model consistently shows superior performance in all feature categories. Notably, combining SWinT with the mel-spectrogram yields accuracy rates from 98.66% to 99.15%. Figure 15a illustrates the confusion matrix corresponding to their peak performance, achieving an average accuracy of 99.02% with an extremely low standard deviation of 0.01%, reflecting efficacy and consistency across all folds. Furthermore, in this configuration, loss scores average out at a mere 0.04, accompanied by a standard deviation less than 0.01, which is indicative of excellent control. Additionally, using chroma-STFT with SWinT produces superior outcomes compared to EffNet, rendering an accuracy of approximately $81.41\% \pm 0.82\%$, and notably exceeding EffNet's $56.46\% \pm 5.05\%$. Meanwhile, MFCC features offer a moderate yet strong and consistent performance with notable accuracy ($\bar{\mathscr{M}} = 95.25\%$ and $\sigma_{\mathscr{M}} = 0.22$), while also maintaining relatively stable loss metrics ($\bar{\mathscr{M}} = 0.17$ and $\sigma_{\mathscr{M}} = 0.01$).

Table 18. Accuracy (in %) and loss scores for MIMII dataset's classification using SWinT

k	Chroma-STFT		MFCC		Mel-spectrogram	
	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss
1	82.15	0.61	95.30	0.17	99.07	0.04
2	80.79	0.64	95.42	0.17	98.66	0.05
3	82.39	0.63	95.46	0.17	99.15	0.04
4	80.56	0.66	95.14	0.16	99.13	0.05
5	81.18	0.63	94.93	0.18	99.11	0.04
$\bar{\mathscr{M}}$	81.41	0.63	95.25	0.17	99.02	0.04
$\sigma_{\mathscr{M}}$	0.82	0.02	0.22	0.01	0.01	0.00



(a) MIMII **(b)** ESC-50 **(c)** FSC22

Fig. 15. Confusion matrix during the highest achievement of MIMII classification with SWinT

Table 19. Accuracy (in %) and loss scores for ESC-50 dataset's classification using SWinT

k	Chroma-STFT		MFCC		Mel-spectrogram	
	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss
1	46.75	2.01	61.00	1.49	84.50	0.51
2	49.75	1.82	59.00	1.53	83.00	0.52
3	50.25	1.70	63.75	1.31	86.75	0.49
4	47.00	1.87	61.75	1.49	81.00	0.66
5	47.25	1.89	60.25	1.37	84.75	0.56
$\bar{\mathscr{M}}$	48.20	1.86	61.15	1.44	84.00	0.55
$\sigma_{\mathscr{M}}$	1.66	0.11	1.77	0.09	2.14	0.07

In the task of categorizing the ESC-50 dataset, the vanilla SWinT model clearly exceeds EffNet in performance, as revealed through its consistent excellence across the three methods of feature extraction, which are thoroughly documented in Tables 16 and 19. In particular, the synthesis of the features mel-spectrogram with SWinT results in a maximum accuracy of 84.00% on average, with Figure 15b

showcasing its superior confusion matrix and a minimal average loss of 0.55. In contrast, when paired with EffNet, this same setup only manages to achieve a mere 59.00% accuracy and suffers from a higher mean loss of 1.49. Similarly, SWinT excels in employing MFCC features by obtaining an average accuracy of 61.15% alongside a decreased mean loss of 1.44, outperforming EffNet's 55.05% accuracy and 1.66 loss. Even with the least effective feature set, chroma-STFT, SWinT still leads with a mean accuracy of 48.20%, which markedly better than EffNet's 27.65%. Analysis of average loss values underscores the robustness of SWinT, demonstrating equivalent or decreased variability for MFCC and mel-spectrogram features (0.09 and 0.07, respectively), in contrast to EffNet's (0.27 and 0.07). In general, these findings confirm that SWinT not only delivers superior classification performance on average, but also provides enhanced reliability across diverse data folds, especially when using perceptually substantial features such as the mel-spectrogram, demonstrating its effectiveness in addressing intricate environmental sound classification issues within ESC-50.

Table 20. Accuracy (in %) and loss scores for FSC22 dataset's classification using SWinT

k	Chroma-STFT		MFCC		Mel-spectrogram	
	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss
1	44.44	1.77	66.91	1.09	76.79	0.75
2	55.06	1.42	66.17	1.20	80.00	0.69
3	55.31	1.42	62.96	1.21	75.06	0.87
4	46.91	1.68	62.72	1.35	73.09	0.87
5	50.37	1.68	62.47	1.44	73.83	0.89
$\mathscr{M}$	50.42	1.59	64.25	1.26	75.75	0.82
$\sigma_{\mathscr{M}}$	4.84	0.16	2.12	0.14	2.76	0.09

Mirroring the outcomes observed with the MIMII and ESC-50 datasets, the SWinT's ability to surpass EffNet is evident in FSC22 classifications as well. An extensive comparison between the SWinT and EffNet models on the FSC22 dataset is presented in Tables 17 and 20, clearly showcasing the superior classification performance of SWinT across three distinct audio feature types. The SWinT architecture, particularly when employing mel-spectrogram features, achieves an impressive mean accuracy of 75.75% and a notably lower mean loss of 0.82, thus consistently outperforming its rival. Conversely, the EffNet model registers only a 48.60% accuracy with a mean loss of 1.77 during implementation with the mel-spectrogram. The confusion matrix for this optimal SWinT configuration is depicted in Figure 15c. Additionally, SWinT exhibits significant improvements in setups based on MFCC, with a mean accuracy of 64.25% and a loss of 1.26, far exceeding EffNet's 53.98% accuracy and 1.74 loss. Even when working with chroma-STFT features, which generally perform less effectively in non-tonal forest sound environments, SWinT still achieves better results, with a mean accuracy of

50.42% and a reduced loss of 1.59, compared to EffNet's 31.16% accuracy and 2.49 loss. The SWinT further demonstrates more consistent performance, as evidenced by lower standard deviations in both accuracy and loss, particularly with MFCC (standard deviation: 2.12% accuracy, 0.14 loss) and mel-spectrogram (standard deviation: 2.76% accuracy, 0.09 loss), highlighting its reliable and steady output over various validation folds. Altogether, these results underscore that SWinT not only significantly enhances the classification accuracy but also offers more generalized learning with reduced overfitting risk. This robustness makes SWinT an ideal candidate for the complex and acoustically varied scenarios embodied in the FSC22 dataset.

3.1.3. Proposed Models

An examination of Tables 15, 21, and 22 highlights pronounced differences in classification performance among the three EffNet-based architectures on the MIMII dataset. Across all models, mel-spectrogram features consistently yield better results than both MFCC and chroma-STFT. Among the architectures, EffNet-BiLSTM generally attains the highest classification accuracy and the lowest loss for nearly all feature configurations. When using the mel-spectrogram inputs, EffNet-BiLSTM reaches the best average accuracy (98.42%) and the lowest mean loss (0.07), coupled with minimal variance ($\sigma_{\mathscr{M}} = 0.31$ and 0.01 for accuracy and loss, respectively), which indicates a highly stable and reliable high-performing model. The EffNet-BiGRU model follows, with a solid average accuracy of 96.37% and a mean loss of 0.14, albeit with noticeably higher variability ($\sigma_{\mathscr{M}}$ values of 1.94 for accuracy and 0.08 for loss). In contrast, the baseline EffNet model, despite reaching a competitive mean accuracy of 97.43%, suffers from substantial variance and elevated loss metrics (3.15 and 6.79 for loss $\bar{\mathscr{M}}$ and $\sigma_{\mathscr{M}}$, respectively). These inflated values stem largely from outlier losses, such as the value 15.29 observed in the fifth fold, which skews the mean and undermines the model's reliability.

When considering classification based on MFCC features, EffNet-BiLSTM once again performs best, yielding an average accuracy of 92.68% and a moderate loss of 0.27, together with low standard deviations of 1.09% for accuracy and 0.02 for loss. By comparison, EffNet-BiGRU attains a slightly lower accuracy of 89.42% and a higher loss of 0.37, though it remains relatively consistent across folds. The vanilla EffNet still manages to be competitive here, achieving 89.54% accuracy and 0.37 loss on average. The results further indicate that chroma-STFT is not well-suited for industrial acoustic data, as it persistently yields inferior performance for all models. EffNet records the lowest mean accuracy at 56.46%, with a standard deviation of 5.05%. EffNet-BiGRU raises the mean accuracy to 65.06%, and EffNet-BiLSTM improves it further to 72.34%. The gradual reduction in accuracy variance across these models suggests that the incorporation of the recurrent layers enhances robustness, even when relying on suboptimal feature representations.

As detailed in Tables 18, 23, and 24, classification results using attention- or

Table 21. Accuracy (in %) and loss scores for MIMII dataset's classification using EffNet-BiGRU

	Chroma-STFT		MFCC		Mel-spectrogram	
k	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss
1	64.29	1.07	89.35	0.38	96.67	0.12
2	65.84	1.03	89.21	0.38	97.68	0.09
3	65.63	1.08	88.76	0.39	97.21	0.11
4	64.33	1.07	90.00	0.36	97.34	0.10
5	65.21	1.05	89.80	0.34	92.97	0.28
$\bar{\mathscr{M}}$	65.06	1.06	89.42	0.37	96.37	0.14
$\sigma_{\mathscr{M}}$	0.72	0.02	0.49	0.02	1.94	0.08

Table 22. Accuracy (in %) and loss scores for MIMII dataset's classification using EffNet-BiLSTM

	Chroma-STFT		MFCC		Mel-spectrogram	
k	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss
1	69.16	0.95	92.06	0.29	97.94	0.08
2	73.85	0.87	94.01	0.24	98.73	0.06
3	72.61	0.86	92.34	0.28	98.56	0.06
4	72.69	0.91	91.40	0.28	98.32	0.07
5	73.38	0.93	93.60	0.25	98.55	0.06
$\bar{\mathscr{M}}$	72.34	0.90	92.68	0.27	98.42	0.07
$\sigma_{\mathscr{M}}$	1.85	0.04	1.09	0.02	0.31	0.01

Table 23. Accuracy (in %) and loss scores for MIMII dataset's classification using SWinT-BiGRU

	Chroma-STFT		MFCC		Mel-spectrogram	
k	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss
1	80.20	0.66	94.89	0.20	98.77	0.05
2	80.70	0.65	95.05	0.18	98.49	0.06
3	79.37	0.68	94.34	0.19	98.79	0.05
4	79.84	0.68	93.80	0.21	98.88	0.05
5	80.07	0.68	94.15	0.21	99.04	0.05
$\bar{\mathscr{M}}$	80.04	0.67	94.45	0.20	98.79	0.05
$\sigma_{\mathscr{M}}$	0.49	0.01	0.52	0.01	0.20	0.00

transformer-based models, namely SWinT, SWinT-BiGRU, and SWinT-BiLSTM, demonstrate superior performance compared to EffNet-based alternatives, particularly when employing mel-spectrogram features. Among these, the basic SWinT model attains the highest mean accuracy of 99.02% with the minimal average loss of 0.04, paired with an impressively low standard deviation (0.01% for accuracy and ≈ 0 for loss), underscoring its impressive consistency and generalization capability.

Table 24. Accuracy (in %) and loss scores for MIMII dataset's classification using SWinT-BiLSTM

	Chroma-STFT		MFCC		Mel-spectrogram	
k	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss
1	79.64	0.66	95.24	0.16	98.94	0.04
2	80.99	0.64	95.15	0.17	98.98	0.04
3	80.92	0.67	94.81	0.18	99.08	0.04
4	79.81	0.68	94.90	0.18	98.84	0.05
5	79.53	0.72	94.48	0.20	99.00	0.04
$\bar{\mathscr{M}}$	80.18	0.67	94.92	0.18	98.97	0.04
$\sigma_{\mathscr{M}}$	0.72	0.03	0.30	0.01	0.09	0.00

Following closely, the SWinT-BiLSTM model reaches an average accuracy of 98.97% and a marginally reduced average loss of 0.04, while preserving low variance. In contrast, the SWinT-BiGRU model shows slightly lower average accuracy at 98.79% and a slightly higher average loss of 0.05, yet remains highly stable across different data partitions. When utilizing MFCCs, all three models exhibit impressive accuracy exceeding 94%, with SWinT excelling at 95.25%, followed by SWinT-BiLSTM at 94.92%, and finally, SWinT-BiGRU at 94.45%. However, the discrepancies are more noticeable as SWinT yields the lowest average MFCC loss (0.17), followed by SWinT-BiLSTM (0.18) and SWinT-BiGRU (0.20). For chroma-STFT, which is generally less beneficial in industrial audio settings, the traditional SWinT surpasses its competitors with an average accuracy of 81.41%, succeeded by SWinT-BiLSTM at 80.18% and SWinT-BiGRU at 80.04%. Ultimately, the basic SWinT model marginally outshines its hybrid counterparts in terms of the general accuracy and reliability when paired with the mel-spectrogram features, while the BiLSTM layers occasionally present competitive and potentially more effective loss reduction when configured with MFCC.

Table 25. Accuracy (in %) and loss scores for ESC-50 dataset's classification using EffNet-BiGRU

	Chroma-STFT		MFCC		Mel-spectrogram	
k	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss
1	30.25	2.46	53.50	1.52	56.00	1.54
2	29.25	2.38	51.25	1.73	58.00	1.40
3	28.50	2.42	49.25	1.88	53.50	1.59
4	29.25	2.44	53.25	1.52	60.00	1.42
5	32.75	2.48	55.25	1.53	61.25	1.46
$\bar{\mathscr{M}}$	30.00	2.44	52.50	1.64	57.75	1.48
$\sigma_{\mathscr{M}}$	1.66	0.04	2.30	0.16	4.70	0.08

Upon analyzing the data in Tables 16, 25, and 26, the significant impact of

Table 26. Accuracy (in %) and loss scores for ESC-50 dataset's classification using EffNet-BiLSTM

	Chroma-STFT		MFCC		Mel-spectrogram	
k	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss
1	31.25	2.28	63.50	1.17	68.25	1.15
2	33.75	2.04	64.25	1.07	62.75	1.18
3	33.00	2.42	69.00	1.02	68.50	1.07
4	34.25	2.13	70.75	0.97	71.25	0.98
5	35.00	2.22	74.25	0.86	59.75	1.46
$\bar{\mathscr{M}}$	33.45	2.22	68.35	1.02	66.10	1.17
$\sigma_{\mathscr{M}}$	1.43	0.14	4.51	0.12	3.10	0.18

adding recurrent layers to the EffNet architecture can be observed when classifying the ESC-50 dataset. Initially, the baseline EffNet model shows limited success, achieving average accuracies of merely 27.65%, 55.05%, and 59.00% when utilizing chroma-STFT, MFCC, and mel-spectrogram, respectively. However, incorporating a recurrent component such as EffNet-BiGRU increases performance only on a selective extractor, with the average accuracies reaching 30.00%, 52.50%, and 57.75%, respectively, in the same order as EffNet implementation. More noteworthy is the configuration of EffNet-BiLSTM, which significantly increases the classification accuracy to 33.45% for chroma-STFT, 68.35% for MFCC, and 66.10% for the mel-spectrogram, illustrating the advancements of 5.80%, 13.3%, and 7.1% over the baseline values. This configuration also demonstrates substantial decreases in loss, notably falling from 1.66 to 1.02 with MFCC, suggesting more efficient convergence and improved generalizability. Moreover, the $\sigma_{\mathscr{M}}$ scores indicate enhanced stability in the EffNet-BiLSTM model in five-fold validations, particularly with the MFCC and the mel-spectrogram, where its variance remains less than or comparable to that of both EffNet and EffNet-BiGRU. These findings verify the importance of embedding recurrent structures such as BiGRU and BiLSTM into the EffNet structure to significantly bolster its ability to discern temporal patterns in environmental sound classification, with EffNet-BiLSTM emerging as the most precise and reliable model in all feature modalities analyzed.

By incorporating recurrent layers before the final classification stage, namely, BiGRU and BiLSTM, vanilla SWinT follows a similar improvement trajectory as observed previously in EffNet for ESC-50 classification. An analysis of Tables 19, 27, and 28 reveals that enhancing the SWinT backbone with BiGRU and BiLSTM recurrent layers significantly improves performance. The baseline SWinT model, leveraging the mel-spectrogram, achieves a mean accuracy of 84.00% with a mean loss of 0.55. Incorporating BiGRU in the SWinT-BiGRU configuration marginally boosts stability and accuracy, attaining averages of 62.75% for MFCC and 84.25% for the mel-spectrogram. The reduction in loss values also reflects smoother convergence. Notably, the integration of BiLSTM in the SWinT-BiLSTM setup delivers the

Table 27. Accuracy (in %) and loss scores for ESC-50 dataset's classification using SWinT-BiGRU

	Chroma-STFT		MFCC		Mel-spectrogram	
k	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss
1	44.00	2.00	69.00	1.24	86.75	0.49
2	47.25	1.73	60.75	1.35	84.75	0.47
3	48.75	1.63	59.00	1.39	85.25	0.43
4	47.50	1.77	61.25	1.32	81.25	0.64
5	46.50	1.73	63.75	1.16	83.25	0.49
$\bar{\mathscr{M}}$	46.80	1.77	62.75	1.29	84.25	0.50
$\sigma_{\mathscr{M}}$	1.76	0.14	3.89	0.09	2.09	0.08

Table 28. Accuracy (in %) and loss scores for ESC-50 dataset's classification using SWinT-BiLSTM

	Chroma-STFT		MFCC		Mel-spectrogram	
k	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss
1	53.50	1.57	71.25	1.01	90.25	0.30
2	56.75	1.34	70.00	0.92	91.50	0.21
3	58.00	1.36	74.50	0.88	93.50	0.22
4	53.25	1.49	70.50	0.91	89.00	0.30
5	52.00	1.54	75.25	0.51	91.50	0.25
$\bar{\mathscr{M}}$	54.70	1.46	72.30	0.85	91.15	0.26
$\sigma_{\mathscr{M}}$	2.55	0.10	2.41	0.19	1.67	0.04

highest results across metrics and features. In conjunction with the mel-spectrogram, SWinT-BiLSTM reaches a peak mean accuracy of 91.15% and a minimal loss of 0.26, outperforming SWinT and SWinT-BiGRU. Furthermore, for MFCC, SWinT-BiLSTM achieves a mean accuracy of 72.30% with a lowered mean loss of 0.85. Additionally, the $\sigma_{\mathscr{M}}$ of accuracy values for SWinT-BiLSTM point to greater stability, especially when the mel-spectrogram is used, producing an accuracy $\sigma_{\mathscr{M}}$ of 1.67%, as opposed to 2.14% and 2.09% for SWinT and SWinT-BiGRU, respectively. These findings emphasize that the incorporation of the temporal modeling through BiLSTM layers significantly enhances the SWinT framework's capacity to capture the sequence characteristics of environmental sounds, improving classification performance and generalization on the ESC-50 dataset.

Taking a similar trajectory, the integration of RNN-based layers into EffNet for the classification of the FSC22 dataset, characteristic of a forest environment, shows notable performance enhancements. By comparing Tables 17, 29, and 30, considerable advances in classification precision are observed after the inclusion of recurrent layers within the EffNet framework. Initially, the baseline model EffNet generates accuracy rates of 31.16%, 53.98%, and 48.60% when using the chroma-STFT, the MFCC, and the mel-spectrogram, respectively, along with loss

Table 29. Accuracy (in %) and loss scores for FSC22 dataset's classification using EffNet-BiGRU

	Chroma-STFT		MFCC		Mel-spectrogram	
k	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss
1	41.23	2.00	62.47	1.24	50.62	1.48
2	34.81	2.14	60.74	1.37	55.31	1.34
3	34.07	2.24	57.78	1.30	51.60	1.47
4	39.51	2.07	57.53	1.33	56.54	1.47
5	28.15	2.32	53.09	1.44	59.26	1.40
$\bar{\mathscr{M}}$	35.55	2.15	58.32	1.33	54.67	1.43
$\sigma_{\mathscr{M}}$	5.13	0.13	3.58	0.07	3.56	0.06

Table 30. Accuracy (in %) and loss scores for FSC22 dataset's classification using EffNet-BiLSTM

	Chroma-STFT		MFCC		Mel-spectrogram	
k	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss
1	49.14	1.73	72.59	0.78	60.74	1.18
2	39.01	1.97	66.67	0.97	68.64	0.99
3	44.69	1.81	81.48	0.57	63.21	1.07
4	37.53	2.03	79.51	0.66	62.22	1.23
5	40.00	2.04	75.31	0.74	55.31	1.36
$\bar{\mathscr{M}}$	42.07	1.92	75.11	0.74	62.02	1.17
$\sigma_{\mathscr{M}}$	4.77	0.14	5.86	0.15	4.79	0.14

values of 2.49, 1.74, and 1.77, in the same order. Enriching EffNet with BiGRU (EffNet-BiGRU) results in noticeable accuracy gains for all types of features, reaching 58.32% for MFCC while reducing the loss to 1.33. Likewise, the EffNet-BiGRU model maintains consistent improvements with the mel-spectrogram (54.67% accuracy and 1.43 loss), although advances in chroma-STFT are less pronounced. However, the structure EffNet-BiLSTM records the highest classification success, especially with MFCC, reaching a mean accuracy of 75.11% and a significantly decreased loss rate of 0.74, surpassing both EffNet and EffNet-BiGRU. For the mel-spectrogram, EffNet-BiLSTM progresses to 62.02% in accuracy and 1.17 in loss, and, with chroma-STFT, it achieves a 42.07% accuracy, exceeding the baseline by approximately 11%. The calculations $\sigma_{\mathscr{M}}$ for EffNet-BiLSTM in various features show stable results, falling within the more restricted limits, notably, for mel-spectrogram (4.79%) and MFCC (5.86%), which are considerably lower than the statistics $\sigma_{\mathscr{M}}$ for EffNet. In general, these results illustrate that the incorporation of temporal modeling layers such as BiGRU and BiLSTM into EffNet substantially increases the model's proficiency in deciphering sequential sound patterns from wilderness audio, with EffNet-BiLSTM showing the best precision and robustness throughout the FSC22 dataset.

Table 31. Accuracy (in %) and loss scores for FSC22 dataset's classification using SWinT-BiGRU

	Chroma-STFT		MFCC		Mel-spectrogram	
k	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss
1	46.67	1.75	66.67	1.02	78.77	0.66
2	56.30	1.44	66.67	1.07	76.79	0.59
3	50.86	1.52	64.20	1.10	78.77	0.70
4	41.48	1.76	64.20	1.15	76.30	0.73
5	44.94	1.78	64.94	1.22	76.79	0.71
$\bar{\mathscr{M}}$	48.05	1.65	65.34	1.11	77.48	0.68
$\sigma_{\mathscr{M}}$	5.72	0.16	1.25	0.08	1.19	0.06

Table 32. Accuracy (in %) and loss scores for FSC22 dataset's classification using SWinT-BiLSTM

	Chroma-STFT		MFCC		Mel-spectrogram	
k	Accuracy (%)	Loss	Accuracy (%)	Loss	Accuracy (%)	Loss
1	44.44	1.77	70.86	0.84	83.70	0.47
2	55.06	1.42	71.11	0.84	82.96	0.49
3	55.31	1.42	72.35	0.83	83.21	0.44
4	46.91	1.68	72.59	0.87	83.46	0.49
5	50.37	1.68	65.93	1.06	82.47	0.51
$\bar{\mathscr{M}}$	50.42	1.59	70.57	0.89	83.16	0.48
$\sigma_{\mathscr{M}}$	4.84	0.16	2.70	0.10	0.47	0.03

Tables 20, 31, and 32 present an in-depth analysis of the effect that the addition of recurrent layers has on the architecture SWinT when classifying the FSC22 dataset. Initially, the baseline model SWinT shows reliable accuracy, particularly when using mel-spectrogram features, achieving an average accuracy of 75.75% and an average loss of 0.82. Incorporation of BiGRU into the SWinT-BiGRU model delivers a slight improvement in performance, especially noted in both MFCC and mel-spectrogram feature spaces. Unlike other datasets, the SWinT-BiGRU records mean the accuracies of 65.34% with MFCC and 77.48% with the mel-spectrogram, each exceeding the baseline SWinT scores. Furthermore, losses decrease, reaching 1.11 and 0.68 for the MFCC and the mel-spectrogram, respectively. The most significant enhancements are observed with the SWinT-BiLSTM configuration that incorporates BiLSTM layers, resulting in a peak performance with 83.16% mean accuracy and a reduced 0.48 average loss for the mel-spectrogram, alongside 70.57% accuracy and 0.89 loss for MFCC. Furthermore, among the configurations, the SWinT-BiLSTM boasts the lowest $\sigma_{\mathscr{M}}$ values, particularly for the mel-spectrogram, which produces $\sigma_{\mathscr{M}}$ = 0.47%, indicating its superior consistency across folds. These observations underscore the substantial advantage of integrating temporal modeling via BiLSTM, greatly improving the sequence learning prowess, which in turn significantly enhances both

accuracy and generalization when dealing with the complex wilderness audio in the FSC22 dataset.

3.2. Ablation Studies

3.2.1. Accuracy

Figure 16 reports a comparative statistical analysis of classification accuracy on the MIMII dataset, obtained with several DL architectures and three different audio feature representations. The blue, green, and red bars correspond to the mean classification performance for the chroma-STFT, the MFCC, and the mel-spectrogram, respectively. The examined models (i.e., EffNet, EffNet-BiGRU, EffNet-BiLSTM, SWinT, SWinT-BiGRU, and SWinT-BiLSTM) were assessed in terms of their capability to capture and generalize intricate acoustic patterns from industrial sound recordings.

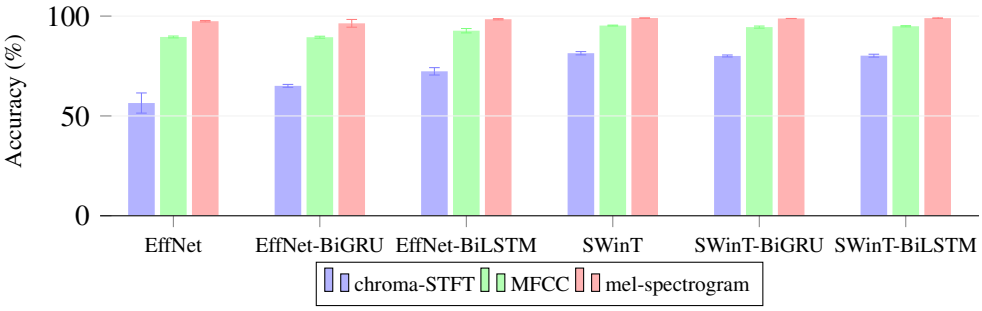


Fig. 16. Comparative statistical analysis of classification accuracies on the MIMII dataset

Among the three features, the chroma-STFT yields the lowest average accuracies, despite being informative for the musical pitch content. Within this feature group, EffNet shows the weakest performance ($56.46\% \pm 5.05\%$), whereas SWinT attains the best result ($81.41\% \pm 0.82\%$), with comparatively narrow confidence intervals indicating more stable predictions. When switching to MFCC representations, the performance improves substantially. Owing to their compact and noise-robust spectral encoding, MFCC features enable a clear accuracy gain, with SWinT and its hybrid variants (SWinT-BiGRU and SWinT-BiLSTM) reaching accuracies of approximately 95%.

The use of the mel-spectrogram further advances performance beyond both chroma-STFT and MFCC across all evaluated models. The best results emerge when mel-spectrogram inputs are paired with transformer-based architectures, where SWinT achieves $99.02\% \pm 0.02\%$, closely followed by SWinT-BiLSTM and SWinT-BiGRU with $98.97\% \pm 0.00\%$ and $98.79\% \pm 0.02\%$, respectively. Additionally, Figure 15a shows the confusion matrix corresponding to the single best-performing experiment on the entire MIMII dataset, which attains an accuracy of

99.13%. This result underlines that the dense time–frequency representation provided by the mel-spectrogram is particularly well-suited to transformer models, which are designed to model long-range temporal dependencies in the data.

The error bars in Figure 16 further highlight the robustness of the various model–feature combinations. Configurations leveraging the mel-spectrogram not only deliver the highest accuracies but also exhibit low variance, indicating reliable and consistent behavior across folds. As outlined earlier, these accuracy statistics arise from a rigorous training and evaluation pipeline (Figure 13), which incorporates k-folds CV together with an early stopping strategy to mitigate overfitting and ensure generalizable performance estimates.

Overall, the results emphasize the beneficial interaction between sophisticated model architectures, particularly SWinT and its sequence-model hybrids, and expressive feature representations like the mel-spectrogram, culminating in state-of-the-art performance for industrial sound classification. At the same time, Figure 16 makes clear that both the intrinsic complexity of the acoustic features and the architectural design of the classifier critically shape the performance, especially under the challenging, noise-contaminated conditions characteristic of the MIMII dataset. Nevertheless, despite being less information-dense than the mel-spectrogram, the MFCC remains an attractive option in scenarios with constrained computational resources, as it offers a favorable trade-off between the accuracy and processing cost.

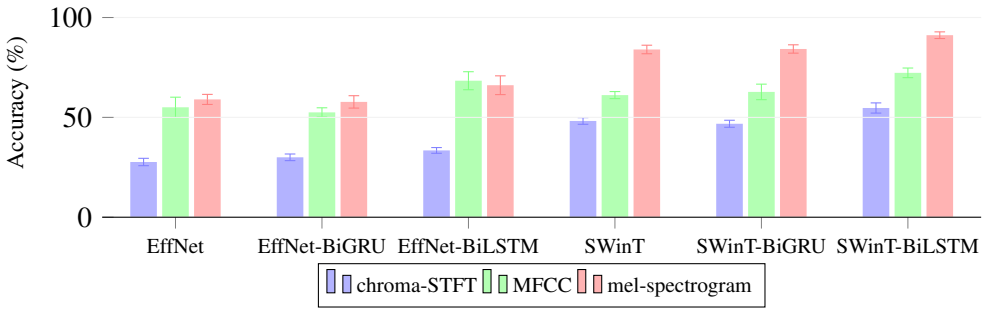
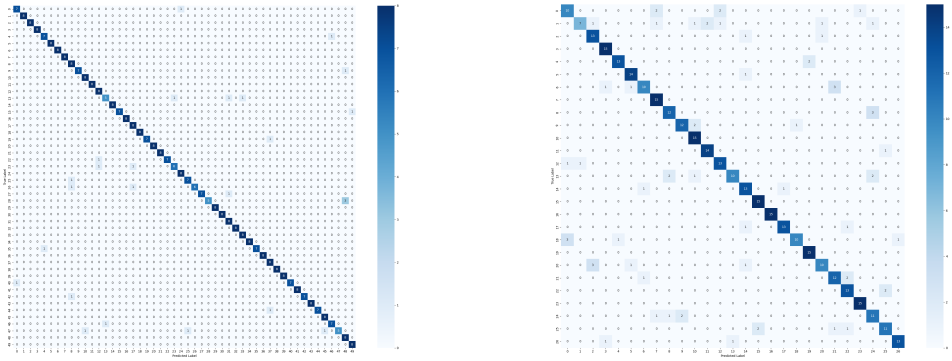


Fig. 17. Comparative statistical analysis of classification accuracies on the ESC-50 dataset

Figure 17 presents a detailed comparison of the classification performance on the ESC-50 dataset, examining six DL architectures: EffNet, EffNet-BiGRU, EffNet-BiLSTM, SWinT, SWinT-BiGRU, and SWinT-BiLSTM. As in earlier works, each model is evaluated in combination with three standard audio feature representations: chroma-STFT, MFCC, and the mel-spectrogram. To improve generalizability and minimize partitioning bias, accuracy is computed via k-folds CV. The results show that the selected audio feature type has a stronger influence on performance than the specific model topology in many cases.

Among all configurations, mel-spectrogram-based features consistently



(a) ESC-50 **(b)** FSC22

Fig. 18. Confusion matrix during the highest achievement of ESC-50 and FSC22 classifications

deliver the best outcomes, with the SWinT-BiLSTM model attaining the highest mean accuracy of $91.15\% \pm 1.67\%$. Complementing these results, Figure 18a illustrates the confusion matrix for this top-performing setup, namely, the mel-spectrogram–SWinT-BiLSTM combination, which achieves an overall accuracy of 91.15%. This visualization further emphasizes the strong discriminative capability of mel-spectrograms for environmental sound recognition, especially when paired with attention-based architectures that effectively exploit long-range temporal information.

In stark contrast, chroma-STFT-derived features yield the weakest performance. For instance, EffNet reaches only $27.65\% \pm 1.85\%$ accuracy, while SWinT-BiLSTM offers a modest improvement to $54.70\% \pm 2.55\%$. These values suggest that chroma-STFT, which focuses primarily on the pitch class information, is ill-suited to capture the heterogeneous and predominantly non-musical events in the ESC-50 dataset. MFCC-based features perform at an intermediate level: they generally surpass chroma-STFT, yet remain clearly below the mel-spectrogram. Their compact, perceptually motivated representation leads to only moderate gains, with the best MFCC configuration—SWinT-BiLSTM—achieving $72.30\% \pm 2.41\%$.

The impact of adding recurrent layers such as BiGRU and BiLSTM to both EffNet and SWinT is nuanced. For MFCC and chroma-STFT inputs, these recurrent components occasionally provide slight accuracy improvements, indicating that explicit sequence modeling can be helpful when the underlying feature representations are less expressive. However, with the mel-spectrogram features, these same recurrent layers often lead to diminished or suboptimal performance. This observation likely arises because attention mechanisms in transformer-based models already capture temporal dependencies effectively, rendering additional recurrent units redundant or even detrimental.

To further ensure reliable model evaluation, training was regularized by an early stopping strategy triggered by stagnation in validation loss. This approach not only mitigates overfitting but also shortens the training time by halting unnecessary epochs. An additional noteworthy result is that configurations based on the mel-spectrogram not only achieve superior mean accuracy but also exhibit very small standard deviations across folds, reflecting high robustness and consistent behavior under k-folds CV.

Taken together, these findings underline the central role of feature expressiveness and the model–feature synergy in environmental sound classification. They demonstrate that rich time–frequency representations, such as mel-spectrograms, when combined with advanced transformer-based architectures like SWinT-BiLSTM, provide the most effective, accurate, and stable framework for addressing the ESC-50 classification task.

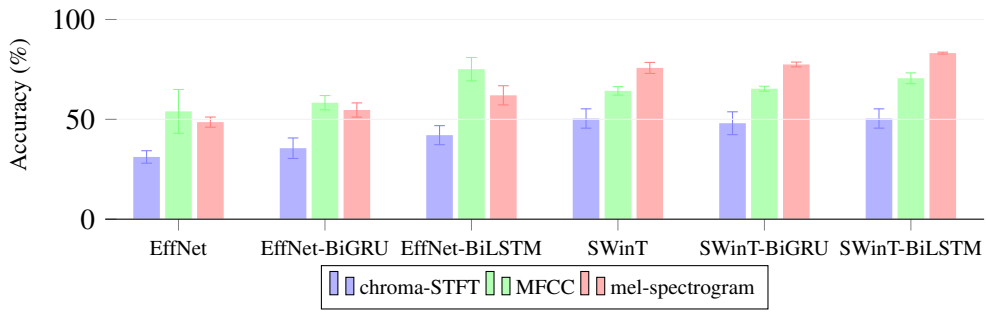


Fig. 19. Comparative statistical analysis of classification accuracies on the FSC22 dataset

Figure 19 presents a comparative overview of the classification accuracy obtained on the FSC22 dataset for six DL architectures, each evaluated in combination with three commonly used audio feature extraction methods. The assessed configurations comprise two base models (EffNet and SWinT) as well as their recurrent-augmented variants that integrate BiGRU and BiLSTM layers on top of the backbone. In the bar plot, blue, green, and red bars denote the mean classification accuracies corresponding to chroma-STFT, MFCC, and mel-spectrogram features, respectively. The vertical error bars indicate the standard deviations estimated through k-folds CV, thereby reflecting the variability across folds. To mitigate overfitting and promote reliable convergence during training, we employed an early-stopping strategy based on monitoring the validation loss and halting optimization once performance ceased to improve.

In line with the findings reported for the MIMII and ESC-50 datasets, mel-spectrogram-based features also yield the highest classification accuracy on FSC22 for almost all models under consideration. Among all evaluated configurations, the pairing of the mel-spectrogram features with the SWinT-BiLSTM architecture

delivers the strongest results, achieving a mean accuracy of $83.16\% \pm 0.47\%$. The confusion matrix corresponding to this best-performing setup, shown in Figure 18b, illustrates clear separation between classes and indicates stable, reliable prediction patterns across the dataset.

In contrast, the use of chroma-STFT features leads to a marked drop in accuracy for all architectures. The baseline EffNet performs worst, achieving only $31.16\% \pm 3.12\%$. The best results with chroma-STFT are obtained by transformer-based models: both SWinT and SWinT-BiLSTM reach similar mean accuracies of about 50.42%, with nearly identical variance. These findings indicate that attention-based mechanisms can only partly compensate for the inherent shortcomings of pitch-class representations. Overall, chroma-STFT appears ill-suited for forest acoustic scene classification, as its representational constraints limit the model's ability to capture the relevant spectral–temporal patterns present in these soundscapes.

The MFCC-based setup yields an intermediate level of performance. Within this feature representation, the recurrent-augmented EffNet-BiLSTM architecture achieves the best overall results, with a mean accuracy of $75.11\% \pm 5.86\%$. It surpasses both the baseline EffNet and all transformer-oriented alternatives (SWinT, SWinT-BiGRU, and SWinT-BiLSTM). This pattern suggests that explicitly modeling temporal dynamics via recurrent layers is especially beneficial when working with compact cepstral descriptors such as MFCCs, likely because these features condense spectral information in a way that makes sequence-dependent processing particularly effective.

Overall, these results highlight the dominant influence of feature representation on the classification accuracy, while also emphasizing the complementary role of architectural design. Rich time–frequency representations such as the mel-spectrogram, when paired with attention-based models like SWinT-BiLSTM, yield the most robust and accurate performance on the FSC22 dataset, whereas recurrent mechanisms offer measurable benefits for lower-dimensional features, such as MFCC.

3.2.2. AUC

Building on the accuracy findings in Figure 16, Figure 20 presents a complementary statistical evaluation based on AUC for MIMII dataset. As argued by Richardson et al. (2024) [181], receiver operation characteristic (ROC)-AUC analysis is particularly appropriate for unbalanced datasets, such as MIMII. Unlike single-threshold metrics like accuracy, AUC offers a threshold-independent measure of how well a model separates normal from abnormal conditions and is less affected by class imbalance.

Across the six DL frameworks and three feature types, a clear trend appears: higher feature complexity yields stronger discriminative performance. As with the accuracy results, chroma-STFT performs the worst, with AUC scores ranging from

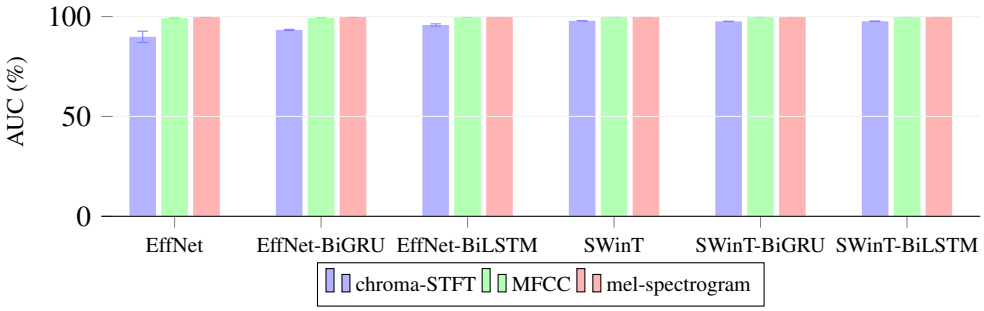


Fig. 20. Comparative statistical analysis of classification AUCs on the MIMII dataset

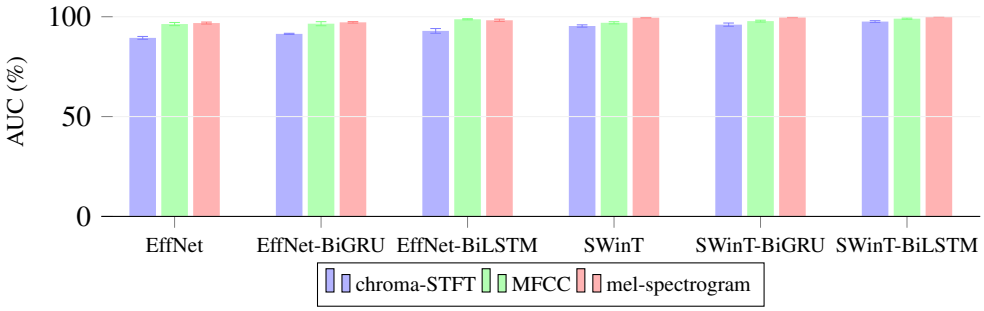


Fig. 21. Comparative statistical analysis of classification AUCs on the ESC-50 dataset

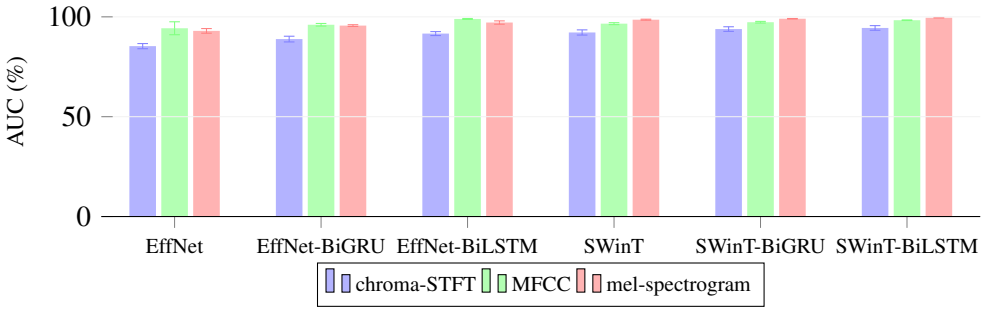


Fig. 22. Comparative statistical analysis of classification AUCs on the FSC22 dataset

89.87% ± 2.83% for EffNet, which also exhibits high variability, to 97.87% ± 0.10% for SWinT. Using MFCC markedly boosts robustness and noise resistance, driving AUC consistently above 99.26%. Performance peaks with the mel-spectrogram, especially when combined with transformer-based models (SWinT, SWinT-BiGRU, SWinT-BiLSTM), where AUC ranges from 99.89% to 99.99%. These transformer variants typically attain 99.99% with minimal cross-fold variation, thereby indicating exceptional stability and reliability.

This extremely consistent behavior, which is in line with the earlier accuracy results, underscores that the mel-spectrogram, owing to their rich time-frequency representation, is particularly effective for capturing the subtle acoustic signatures of faulty industrial machinery. The joint observation of high AUC and low variance further confirms the robustness of these configurations across CV folds. As with the accuracy analysis, all AUC scores were obtained under a uniform k-folds CV protocol with early stopping, supporting the reliability of the reported outcomes. Together with the evidence in Figure 16, these results strengthen the conclusion that advanced feature representations coupled with transformer-based DL architectures are key to building dependable, high-performance audio classification systems for industrial environments.

Figure 21 extends these observations by assessing class-wise discriminability using the AUC metric. We apply the same k-folds CV protocol, partitioning the dataset into equal folds to repeatedly train and validate models, ensuring statistically robust and split-independent results. To curb overfitting and reduce the training time, we adopt early stopping, by terminating training once validation loss fails to improve after a fixed number of epochs. Among the six models, mel-spectrogram-based systems achieve the best AUC scores, with the SWinT-BiLSTM configuration reaching $99.93\% \pm 0.03\%$. This setup not only yields the highest average accuracy in Figure 17, but also delivers strong class separation, which is most evident in architectures that merge attention mechanisms with explicit temporal modeling, such as SWinT-BiLSTM, which sharpen decision boundaries in the learned space. In contrast, chroma-STFT features perform worst: EffNet attains the lowest AUC at $89.40\% \pm 0.71\%$, and even the best chroma-STFT model (SWinT-BiLSTM) only reaches $97.57\% \pm 0.45\%$. MFCC-based systems show intermediate, stable performance, peaking at $99.03\% \pm 0.31\%$ with SWinT-BiLSTM. The narrow spread of AUC values across mel-spectrogram-based models further emphasizes their robustness to different splits. Overall, these AUC results corroborate the accuracy findings and underline the importance of expressive feature representations and transformer-style architectures for building accurate, consistent environmental sound classifiers. Jointly, Figures 17 and 21 indicate that state-of-the-art ESC-50 performance stems from coupling high-resolution audio features with advanced, attention-driven models.

Figure 22 complements the accuracy analysis by reporting the AUC values for all model–feature combinations on FSC22. Consistent with Figure 19, models using mel-spectrogram features clearly outperform those based on MFCC or chroma-STFT. The best AUC is obtained by SWinT-BiLSTM with the mel-spectrogram, reaching $99.55\% \pm 0.04\%$, indicating excellent class separability. Even the basic SWinT combined with the mel-spectrogram attains a strong AUC of $98.57\% \pm 0.26\%$, underscoring the robustness of this feature representation. In contrast, chroma-STFT-based models yield noticeably lower AUC scores, namely, $85.33\% \pm 1.29\%$ for EffNet and $94.44\% \pm 1.16\%$ for SWinT-BiLSTM. As before,

MFCC inputs lead to intermediate performance. In this case, EffNet-BiLSTM achieves $98.95\% \pm 0.25\%$, clearly outperforming the baseline EffNet at $94.29\% \pm 3.24\%$, showing that explicit temporal sequence modeling is especially beneficial for compact cepstral features. As with accuracy, all AUC values were obtained by using k-folds CV with early stopping, thus supporting the reliability and generalizability of the results. Together with the accuracy data, Figure 22 emphasizes the central role of rich feature representations and carefully tuned architectures in solving challenging audio classification tasks.

3.2.3. Precision

Figure 23 displays a comparison of the classification precision of various audio features and DL models using the MIMII dataset. Consistent with the accuracy and AUC analyses, precision was evaluated by using three feature extractors, six DL models, k-folds CV with $k = 5$, and early stopping, following the workflow depicted in Figure 13. According to the figure, the mel-spectrogram consistently demonstrated the highest precision, particularly with SWinT and SWinT-BiLSTM, achieving $99.03\% \pm 0.20\%$ and $98.98\% \pm 0.08\%$, respectively, highlighting its ability to capture time-frequency dynamics which are crucial for detecting complex acoustic anomalies in industry settings. In contrast, the chroma-STFT features offered the lowest precision, especially with the EffNet scoring $62.75\% \pm 3.51\%$, indicating limited efficacy for noisy non-tonal machine sounds. However, its combination with SWinT and its variants were still able to offer scores above 80% consistently. Meanwhile, MFCC showed moderate and reliable precision, with SWinT reaching $95.32\% \pm 0.17\%$, showing that compact features work effectively with advanced models. The achievements also inform that the models with BiGRU or BiLSTM layers performed better than those without them, notably for EffNet variants. Augmenting the attention-based SWinT backbone with recurrent layers, as realized in SWinT-BiGRU and SWinT-BiLSTM, leads to additional gains in precision and, in multiple experimental configurations, especially when using mel-spectrogram inputs, also tends to decrease performance variability across cross-validation folds. Therefore, integrating mel-spectrogram features with attention-driven RNN-enriched models, particularly those based on SWinT, provides the most precise and stable classification of industrial sound anomalies.

Figure 24 shows the precision scores from the use of various audio features and DL models on the ESC-50 dataset. These results corroborate previous analysis and are also derived from the methodology depicted in Figure 13. The figure underlines that employing the mel-spectrogram, alongside the SWinT models, whether in their original or RNN-enhanced forms, has yielded exceptional performance, consistently ranking them among the top three configurations. The pinnacle of performance was achieved through the integration of the mel-spectrogram and SWinT-BiLSTM, which secured a mean precision score of $92.12\% \pm 1.66\%$. This was followed by combinations of the mel-spectrogram with SWinT-BiGRU and SWinT,

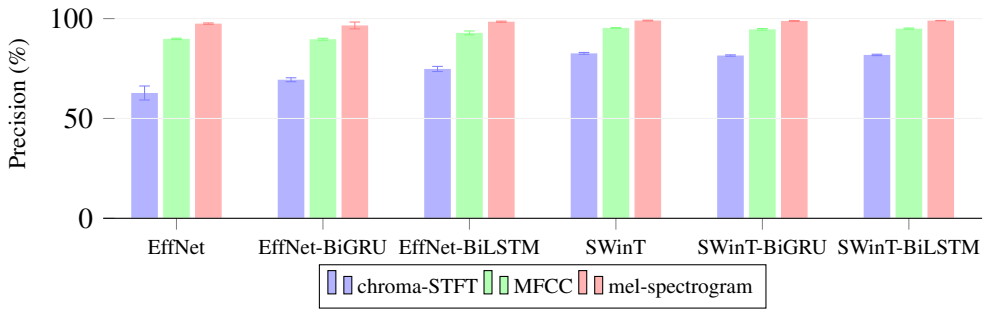


Fig. 23. Comparative statistical analysis of classification precisions on the MIMII dataset

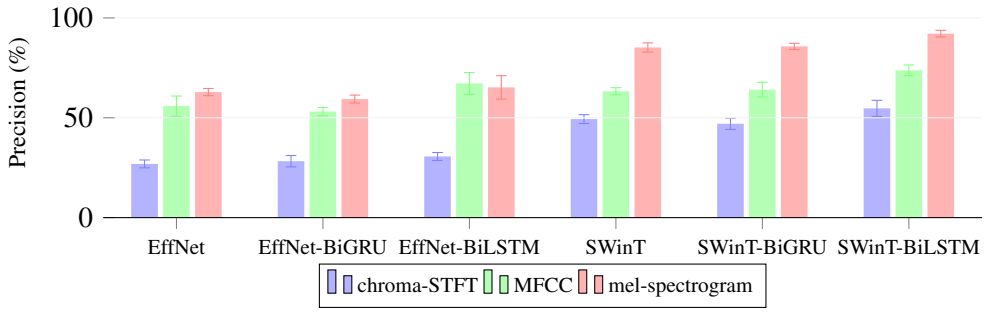


Fig. 24. Comparative statistical analysis of classification precisions on the ESC-50 dataset

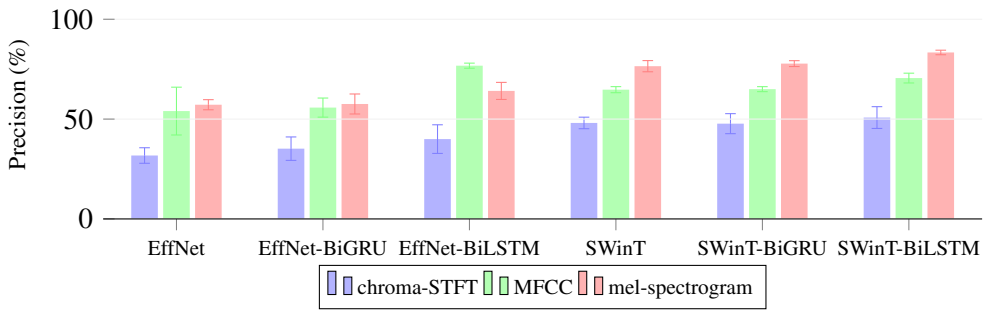


Fig. 25. Comparative statistical analysis of classification precisions on the FSC22 dataset

which obtained mean precision scores of $85.76\% \pm 1.52\%$ and $85.19\% \pm 2.31\%$, respectively. In contrast, all setups involving the chroma-STFT feature extractor underperformed compared to its alternatives, with precision scores that only ranged

from $26.90\% \pm 1.97\%$ (coupled with EffNet) to $54.71\% \pm 4.03\%$ (in association with SWinT-BiLSTM). The MFCC-based feature representation achieves a mid-range level of precision, with performance spanning from $53.12\% \pm 2.03\%$ when using the EffNet-BiGRU model up to $73.75\% \pm 2.70\%$ for the SWinT-BiLSTM architecture. The further analysis in Figure 24 delineates the disparities in the classification properties between ESC-50 and MIMII, particularly in the deployment of BiGRU and BiLSTM layers as enhancements to the foundational models. A comparative examination of Figure 23 and Figure 24 reveals that the scale of score enhancement resulting from enrichment is consistently observed in nearly all feature extraction techniques, extending beyond just the low and medium features seen in MIMII. However, despite this incremental advancement, the classification precision with MFCC and, notably, with chroma-STFT remains unsatisfactory. In contrast to the accuracy patterns observed in Figure 16, the precision obtained by using chroma-STFT on ESC-50 is notably poor for the EffNet family, consistently staying far below the 50% threshold across the evaluated configurations.

Figure 25 provides a comprehensive precision analysis that highlights the influence of the feature depth and model architecture on various models applied to the FSC22 dataset. In particular, integrating the SWinT-BiLSTM model with the features of the mel-spectrogram yielded a precision of $83.36\% \pm 1.14\%$. In contrast, the EffNet model, combined with the chroma-STFT features, generated the lowest precision at $31.76\% \pm 3.89\%$, indicating a significant performance disparity. Regarding the utility of features, the mel-spectrogram outperformed both MFCC and chroma-STFT significantly in all the models evaluated. Its superior performance is due to advanced time-frequency resolution, efficiently capturing the intricate details typical of natural soundscapes. On the other hand, MFCCs exhibited intermediate precision scores, which then improved markedly when complemented with layers of BiGRU or BiLSTM. For example, EffNet-BiLSTM managed to achieve a precision rate of $76.74\% \pm 1.28\%$ with MFCC. This achievement is even better than the products of classification of the combination of MFCC with SWinT or SWinT-BiGRU. In contrast, chroma-STFT produced the least impressive results in all models, which indicates its limitations in handling the non-musical and non-tonal acoustic characteristics common in forest sound recordings. In general, the inclusion of recurrent layers improved the performance, particularly for the features of MFCC and mel-spectrogram, as these layers are proficient in capturing temporal dependencies. In transformer-based models, supplementing with BiLSTM layers, such as in SWinT-BiLSTM, further elevated precision levels, suggesting that combining recurrence with attention mechanisms can be synergistic. In particular, configurations that combine the mel-spectrogram with more advanced modeling approaches, as well as several transformer-based setups, show relatively low variance in their results. The relatively limited variation observed across several of the stronger configurations suggests that these models remain stable and reliable, even when the training–validation split is changed, indicating a robust performance across

different data partitionings. Collectively, these findings underscore the importance of a strategic combination of detailed spectral characteristics and advanced architectural designs that are adept at modeling temporal dynamics for the effective classification of complex environmental audio recordings.

3.2.4. Recall

Another key metric for assessing classification performance on the MIMII dataset is recall (also known as sensitivity). Figure 26 presents a comparison of recall values for six DL models combined with three different audio feature types. Among all model–feature combinations, the mel-spectrogram representation consistently achieves the highest recall, ranging from $96.37\% \pm 1.94\%$ with EffNet-BiGRU to $99.02\% \pm 0.20\%$ with SWinT. This consistent advantage suggests that the rich, densely sampled time–frequency representation provided by mel-spectrograms is particularly effective for encoding fine-grained indicators of machine condition, especially when the recordings are affected by background noise or other forms of contamination in realistic operating environments.

While the MFCC offers a more compact representation than the mel-spectrogram, it still delivers consistently high recall (exceeding 89% across all settings), with a maximum of $95.25\% \pm 0.22\%$ when combined with SWinT. This underscores the practical value of MFCC as a computationally efficient option under limited-resource conditions. By comparison, chroma-STFT, which is mainly tailored to capture the pitch-class structure, performs poorly on industrial machine sounds: the lowest recall occurs with EffNet at $56.46\% \pm 5.05\%$, and even its best-performing setup (SWinT) attains only $81.41\% \pm 0.82\%$, remaining clearly below both MFCC and the mel-spectrogram.

Figure 26 further indicates that adding recurrent layers (GRU/LSTM) on top of the EffNet backbone boosts recall for the lower-capacity feature sets (chroma-STFT and MFCC), aligning with the known advantages of temporal modeling for periodic machine behavior. In contrast, applying a similar recurrent extension to the SWinT backbone does not provide additional gains, as recall is already close to saturated with the mel-spectrogram. Collectively, these findings reinforce the mel-spectrogram as the most effective representation for maximizing recall on MIMII, while MFCC offers a strong balance between performance and computational efficiency.

The trends seen for accuracy, AUC, and precision on ESC-50 are likewise evident in the recall scores presented in Figure 27. Configurations using the mel-spectrogram features consistently surpass the others, with SWinT-BiLSTM achieving the highest mean recall of $91.15\% \pm 1.67\%$. The next best mel-spectrogram-based performances come from SWinT-BiGRU and SWinT, which obtain $84.25\% \pm 2.09\%$ and $84.00\% \pm 2.14\%$, respectively. By contrast, the lowest recall is observed for chroma-STFT with EffNet at $27.65\% \pm 1.85\%$, underscoring that pitch-class representations are poorly matched to the diverse, predominantly non-musical sound events in ESC-50.

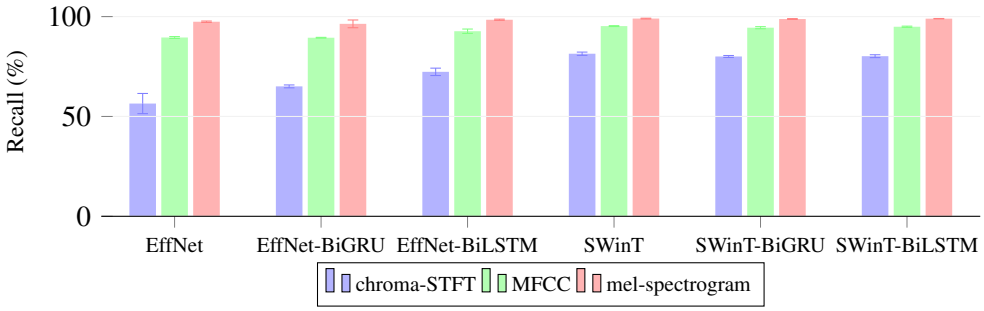


Fig. 26. Comparative statistical analysis of classification recalls on the MIMII dataset

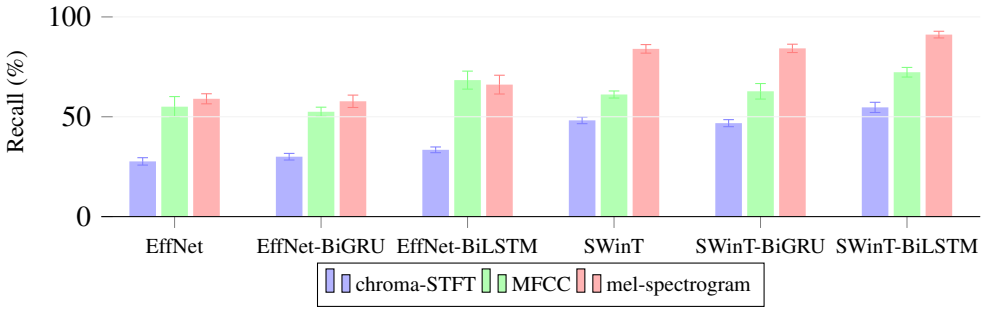


Fig. 27. Comparative statistical analysis of classification recalls on the ESC-50 dataset

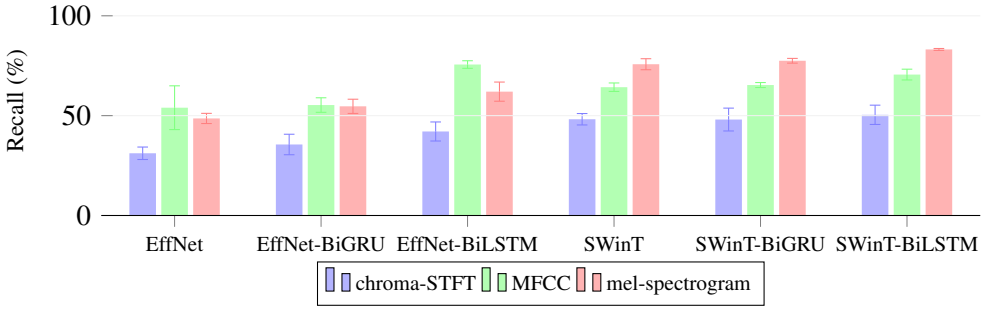


Fig. 28. Comparative statistical analysis of classification recalls on the FSC22 dataset

A marked exception arises for the EffNet family with MFCC features: EffNet-BiLSTM attains $68.35\% \pm 4.51\%$ recall, surpassing the mel-spectrogram-based recall achieved by the same backbone (EffNet-BiLSTM at $66.10\% \pm 4.70\%$). In addition, this MFCC+EffNet-BiLSTM setup outperforms the corresponding MFCC results obtained with SWinT ($61.15\% \pm 1.77\%$) and SWinT-BiGRU ($62.75\% \pm 3.89\%$). These findings suggest that, when using compact cepstral

descriptors, explicitly modeling temporal dynamics through an LSTM-augmented backbone can be especially advantageous. Taken together, Figure 27 reinforces the overall superiority of the mel-spectrogram features for recall on ESC-50, while also illustrating that recurrent augmentation can substantially improve recall for specific feature–backbone combinations within the EffNet family.

Figure 28 presents the recall performance of the examined hybrid DL architectures on the FSC22 dataset [7]. In line with the previous metrics, mel-spectrogram-based representations deliver the highest recall across all architectures. The top-performing setup is SWinT-BiLSTM with the mel-spectrogram, reaching $83.16\% \pm 0.47\%$, while the lowest recall is obtained by chroma-STFT with EffNet at $31.16\% \pm 3.12\%$. These findings further confirm the strong compatibility between dense time-frequency encodings and SWinT-based backbones enhanced with recurrent dynamics modeling.

Across both backbone families, the LSTM-augmented models consistently achieve higher recall than their GRU-based counterparts (for instance, SWinT-BiLSTM surpasses SWinT-BiGRU when using chroma-STFT, MFCC, and mel-spectrogram inputs). This gap is especially large for MFCC, where EffNet-BiLSTM attains $75.60\% \pm 1.90\%$ recall versus $55.31\% \pm 3.64\%$ for EffNet-BiGRU. In contrast, chroma-STFT-driven configurations remain the least effective overall, particularly without recurrent augmentation, suggesting that pitch-centric features alone cannot adequately represent the complex, non-tonal acoustic patterns typical of forest soundscapes. Overall, Figure 28 indicates that transformer–recurrent hybrids paired with the mel-spectrogram inputs offer the most dependable path to high recall on FSC22.

3.2.5. F1

The assessment of F1 scores provides a deeper insight into the classification of the MIMII dataset. Figure 29 visualizes a comparative analysis of the F1 rates for different feature extractors and DL model applications in this study. As the graph suggests, implementations using the mel-spectrogram steadily produce the highest classification quality, regardless of the model used. The mean F1 scores obtained with the the mel-spectrogram range from $96.43\% \pm 1.88\%$ (EffNet-BiGRU) to $99.02\% \pm 0.20\%$ (SWinT). In contrast, as with the other metrics, the utilization of chroma-STFT in MIMII classification generates the least desirable F1 scores, strengthening the indication of its limited suitability for non-musical and non-tonal classification problems. The lowest mean F1 score is obtained from the combination of chroma-STFT and EffNet, yielding only $59.95\% \pm 4.63\%$. Moreover, even the best chroma-STFT result, $81.80\% \pm 0.66\%$ (with SWinT), remains below the weakest outcomes achieved by the higher-capacity feature extractors. Meanwhile, MFCC offers moderate yet reliable classification quality, with all configurations exceeding 89% and peaking at $95.27\% \pm 0.20\%$ when paired with SWinT. This finding, together with the previous metrics, suggests that MFCC remains a viable

option for MIMII classification, particularly under limited computational resources. Additionally, Figure 29 shows that augmenting EffNet with recurrent layers (GRU/LSTM) improves F1 for chroma-STFT and MFCC; however, for SWinT, recurrent augmentation does not yield further gains and can slightly reduce the F1 score in near-saturation conditions.

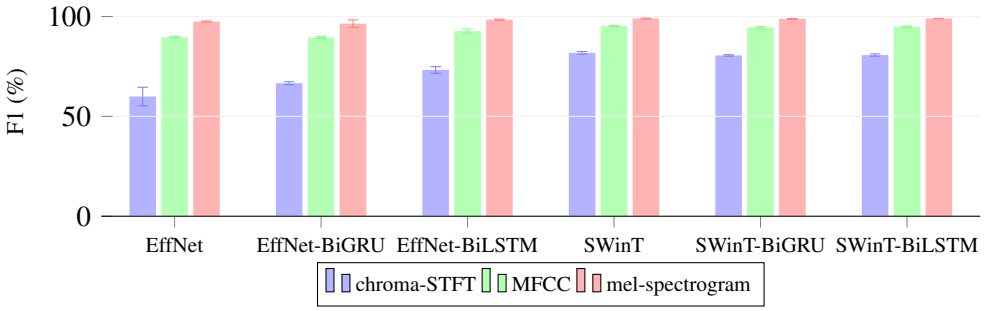


Fig. 29. Comparative statistical analysis of classification F1s on the MIMII dataset

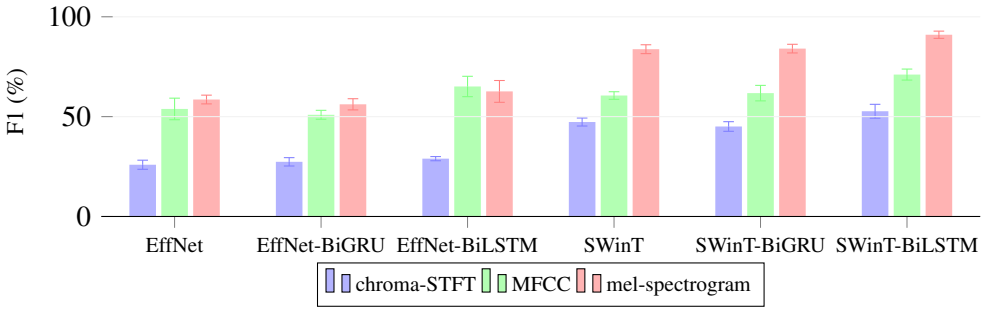


Fig. 30. Comparative statistical analysis of classification F1s on the ESC-50 dataset

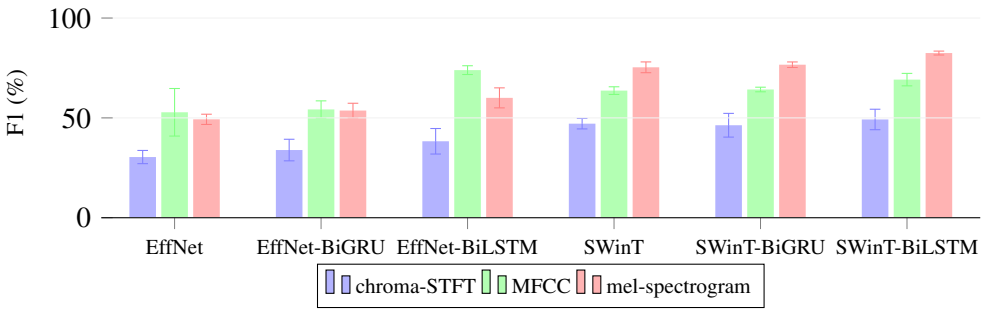


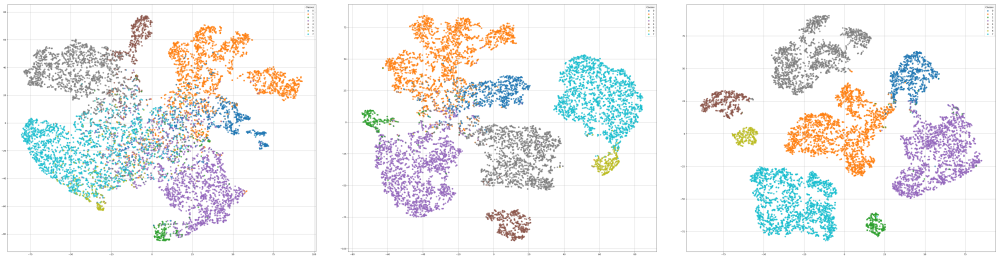
Fig. 31. Comparative statistical analysis of classification F1s on the FSC22 dataset

Conversely, Figure 30 presents a detailed assessment of the F1 scores for

the different model–feature combinations on the ESC-50 dataset. When interpreted together with Figures 17, 21, 24, and 27, similar patterns emerge. Specifically, SWinT, SWinT-BiGRU, and SWinT-BiLSTM gain substantial performance improvements from high-resolution spectral inputs based on the mel-spectrogram, yielding the highest F1 scores, with SWinT-BiLSTM reaching a peak value of $91.01\% \pm 1.81\%$. The SWinT-BiGRU and baseline SWinT variants also perform competitively, achieving $84.06\% \pm 2.17\%$ and $83.77\% \pm 2.22\%$, respectively. By comparison, chroma-STFT persistently delivers the lowest scores across all architectures, while MFCC attains F1 results that lie between those of the mel-spectrogram and chroma-STFT.

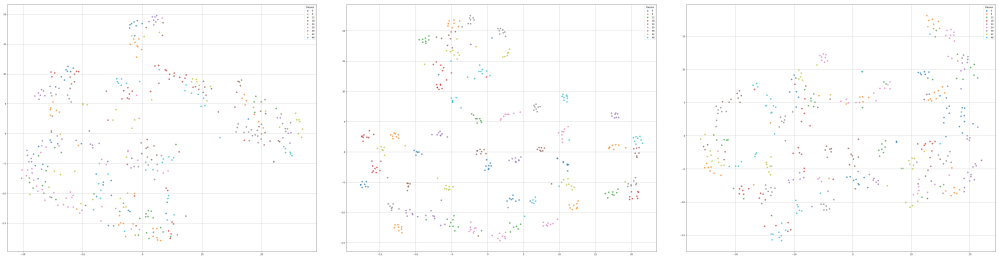
Figure 30 further exposes a clear performance gap within the EffNet family. with the mel-spectrogram features, EffNet-based models perform relatively poorly, with F1 scores between $56.17\% \pm 2.80\%$ and $62.63\% \pm 5.46\%$, and their performance drops even more when chroma-STFT is used. In contrast, MFCC features offer modest gains over chroma-STFT across all architectures, with EffNet variants generally benefiting more than SWinT variants under the same feature configuration. In particular, combining MFCC with EffNet-BiLSTM yields $65.10\% \pm 5.10\%$, surpassing the MFCC-based results of SWinT and SWinT-BiGRU ($60.57\% \pm 1.90\%$ and $61.77\% \pm 3.88\%$, respectively). Overall, the figure emphasizes the decisive role of the feature choice (with the mel-spectrogram performing best, MFCC intermediate, and chroma-STFT worst) and model architecture, consistently showing that SWinT variants outperform EffNet variants in terms of the F1 score on the ESC-50 task.

Figure 31 presents a comparison of the F1 scores for classification on the FSC22 dataset, supporting the trends already observed in Figures 19 to 28. The combination of the mel-spectrogram with SWinT-BiLSTM delivers the best F1 performance on FSC22, achieving $82.47\% \pm 0.98\%$ and demonstrating both high accuracy and low variability across folds. At the opposite end, the use of chroma-STFT yields the poorest results, with F1 scores between $30.39\% \pm 3.31\%$ and $49.21\% \pm 5.13\%$, thereby confirming that pitch-based features are ill-suited to the complex, largely non-tonal nature of forest sound events. Positioned between these extremes, MFCC offers a compromise between the computational cost and descriptive power. The most effective MFCC-based setup combines it with SWinT-BiLSTM, reaching $69.17\% \pm 3.12\%$ under 5-fold CV, whereas the least effective is EffNet with $52.81\% \pm 11.91\%$, which also shows the greatest variability. Overall, these outcomes underscore the importance of recurrent augmentation (GRU/LSTM) for modeling the temporal structure in FSC22 sound events, with LSTM-augmented models generally achieving higher and more consistent F1 scores than their GRU-augmented counterparts for the same feature sets, at the expense of increased computational complexity.



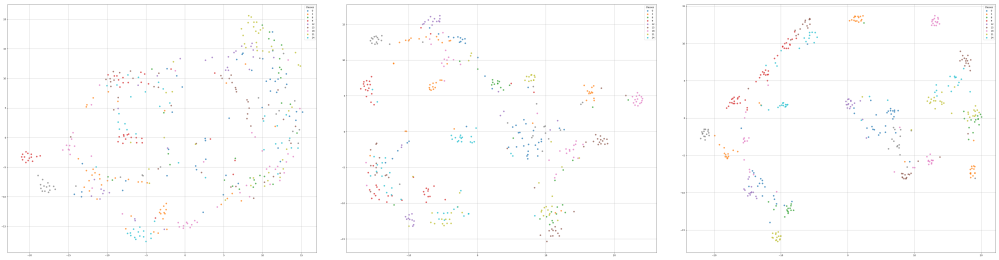
(a) chroma-STFT **(b)** MFCC **(c)** mel-spectrogram

Fig. 32. t-SNE visualization of the highest achievement using distint features in classifying the MIMII dataset



(a) chroma-STFT **(b)** MFCC **(c)** mel-spectrogram

Fig. 33. t-SNE visualization of the highest achievement using distinct features in classifying ESC-50 dataset



(a) chroma-STFT **(b)** MFCC **(c)** mel-spectrogram

Fig. 34. t-SNE visualization of the highest achievement using distinct features in classifying the FSC22 dataset

3.2.6. t-SNE

A comprehensive evaluation of the classification capabilities for the MIMII, ESC-50, and FSC22 datasets requires visualization-driven techniques to complement metric-based assessments. In this regard, the use of analysis t-SNE is particularly

valuable, as it allows a clear depiction of the distributions of high-dimensional audio features in a two-dimensional space [48, 182]. Such visualizations provide insight into the clustering behavior and separability of audio classes, allowing for a more intuitive understanding of model performance across different feature representations. When applying t-SNE, it becomes possible to visually verify the degree of discrimination achieved by various models and characteristics, and to detect patterns that may not be easily captured by precision, loss, or other scalar evaluation metrics alone.

The t-SNE visualizations of the MIMII dataset in Figure 32, generated from three audio feature types (i.e., chroma-STFT, MFCC, and mel-spectrogram), reveal how each representation forms low-dimensional embeddings of machine sounds with differing effectiveness. In Figure 32a, the chroma-STFT-based plot exhibits weakly defined, strongly overlapping clusters with minimal separation, indicating that chroma-STFT is ill-suited for industrial audio. Since it emphasizes pitch classes and harmonic content, chroma-STFT is mismatched to the largely non-tonal, harmony-poor nature of machine acoustics, limiting its usefulness for anomaly-focused condition monitoring. In contrast, Figure 32b shows that MFCC yields more coherent, compact clusters with moderate separation. MFCC better reflects the spectral and temporal traits of mechanical systems, such as the spectral slope, formant structure, and energy distribution, supporting more reliable discrimination between normal and faulty states in fault detection and classification tasks. The clearest structure arises from the mel-spectrogram-based visualization in Figure 32c, which produces well-separated, cohesive clusters with limited overlap. This improved separability stems from the richer time–frequency detail of the mel-spectrogram, enabling accurate capture of temporal and spectral energy patterns that define machine sound signatures. Consequently, mel-spectrogram-derived features provide especially effective input spaces for high-performing DL models in audio anomaly detection, where both temporal evolution and frequency content matter. Overall, the results indicate that mel-spectrogram features are best-suited for MIMII machine condition classification, followed by MFCC, whereas chroma-STFT does not adequately model non-harmonic industrial sound. This comparison highlights the importance of domain-appropriate feature selection and shows how suitable representations improve cluster separability and interpretability in embedded spaces, which is crucial for practical machine monitoring.

The ESC-50 visualizations in Figure 33, based on t-SNE embeddings, are shown for three feature types in Figures 33a (chroma-STFT), 33b (MFCC), and 33c (mel-spectrogram). They reveal marked differences in class separability and preservation of the feature-space structure. Figure 33a shows chroma-STFT embeddings forming diffuse, interwoven clusters with an extensive class overlap, reflecting their weak discriminative power for environmental sounds. This stems from chroma's emphasis on the pitch and harmonic content, making it more suitable for music than the diverse sounds in ESC-50. By contrast, Figure 33b presents MFCC-based embeddings that yield more coherent clusters and better

class distinction, though some boundaries remain ambiguous, consistent with MFCCs offering a perceptually grounded, compact spectral representation. The mel-spectrogram embeddings in Figure 33c exhibit the most clearly defined, compact clusters with well-separated classes, capturing richer temporal–spectral detail that benefits classification. Overall, these comparisons indicate that the mel-spectrogram features provide the strongest class discrimination on ESC-50, followed by MFCCs, with chroma-STFT performing worst for broad environmental sound classification.

The t-SNE visualizations of the FSC22 dataset in Figure 34, generated from three types of audio features, illustrate clear differences in how effectively each representation forms a discriminative low-dimensional space for complex forest sound classification. In Figure 34a, the chroma-STFT projection exhibits overlapping, diffuse clusters, evidencing poor class separation. As a predominantly harmonic descriptor, chroma-STFT fails to capture the non-tonal, transient components typical of forest soundscapes—such as footsteps or animal calls—making it ill-suited for robust categorization. By comparison, Figure 34b shows that MFCC yields a more coherent feature space, with tighter intraclass groupings and improved separation, supported by its psychoacoustic grounding and sensitivity to spectral envelope characteristics. While some overlap persists, MFCC better accommodates the heterogeneous content of FSC22. Most notably, the mel-spectrogram-based projection in Figure 34c presents well-defined clusters and the strongest class separation, capturing the rich time–frequency structure of natural soundscapes. This clear spatial organization indicates that the mel-spectrogram provides a robust feature representation, well-suited for training DL models with high classification performance in diverse acoustic conditions. Overall, mel-spectrogram emerges as the most effective representation for FSC22, with MFCC offering intermediate performance and chroma-STFT performing poorly due to its mismatch with environmental audio properties. These results underscore the critical role of the feature choice in audio classification, particularly in complex acoustic environments that demand detailed spectral–temporal modeling for accurate event recognition.

3.2.7. Vulnerability to Overfitting

This research not only assesses typical performance indicators such as accuracy, precision, recall, and the F1 score, but also investigates training and validation patterns for models using the MIMII, ESC-50, and FSC22 datasets. Special emphasis is placed on the epoch count necessary before engaging the early stopping mechanism, which prevents further training when improvements stagnate, thus avoiding overfitting. Analyzing the behavior of the model during training offers crucial insight into how the different features of each dataset influence learning effectiveness and generalization. The MIMII dataset is notably imbalanced, with a majority of normal operating condition recordings compared to fault samples, potentially leading to skewed learning and heightened overfitting risk if training extends past peak generalization [4]. On the other hand, ESC-50 and FSC22 feature more balanced class

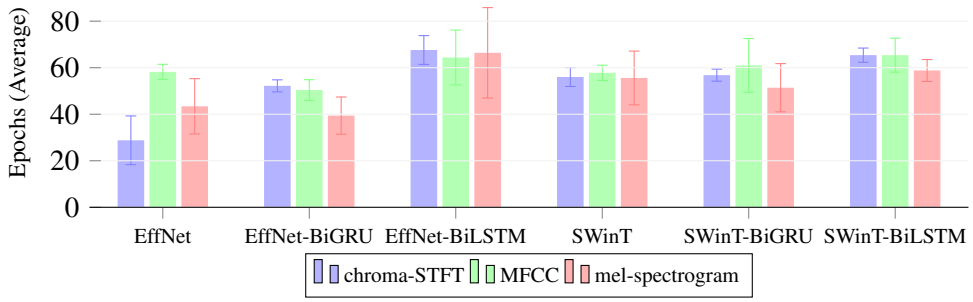


Fig. 35. Comparative statistical analysis of the number of classification epochs on the MIMII dataset

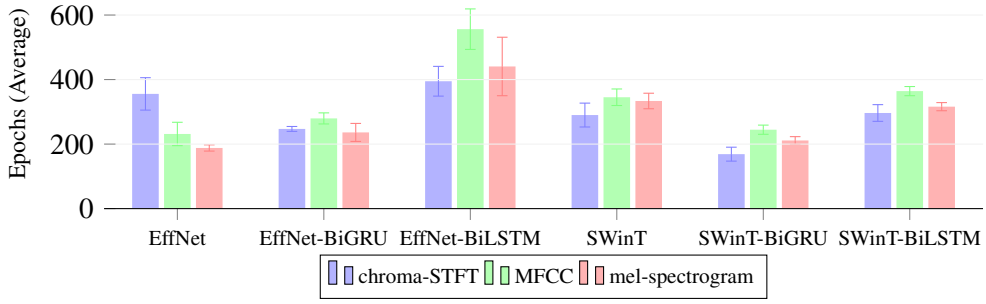


Fig. 36. Comparative statistical analysis of the number of classification epochs on the ESC-50 dataset

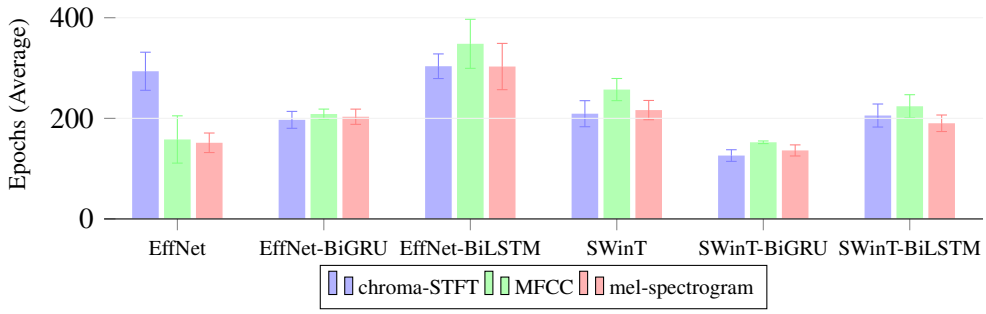


Fig. 37. Comparative statistical analysis of the number of classification epochs on the FSC22 dataset

distributions but vary greatly in content; ESC-50 includes controlled environmental sounds such as animal calls and human activities [5], while FSC22 captures complex forest acoustics, characterized by event overlaps and unpredictable noise [7]. These

distinctions affect the epochs needed; more intricate datasets may require prolonged training to assimilate detailed patterns, though this also elevates memorization risks if early stopping is mismanaged. Hence, careful monitoring of the duration of the training is vital to avoid overfitting and to ensure that the model maintains a strong generalization on new data [131].

Figure 35 provides a detailed comparative analysis of the average number of training epochs required for different model architectures and input feature representations on the MIMII dataset. The horizontal axis categorizes six hybrid models: EffNet, EffNet-BiGRU, EffNet-BiLSTM, SWinT, SWinT-BiGRU, and SWinT-BiLSTM, which combine basic EffNet and SWinT backbones with variants augmented by using GRU and LSTM recurrent layers. Each model group features color-coded bars that represent the averages of the chroma-STFT, MFCC, and mel-spectrogram characteristics, accompanied by error bars indicating standard deviation. The figure generally illustrates that the imbalanced characteristics of the dataset, as highlighted in Table 5, tend to terminate training prematurely, regardless of the selection of the model or feature. Although the maximum number of training epochs is set at 2000, as detailed in Table 12, early stopping frequently intervened, preventing any combination of models and features from reaching the 100th epoch. The highest mean epoch count is observed for the EffNet-BiLSTM model with the mel-spectrogram, averaging 66.4 epochs, while the longest individual run across folds reaches 97 epochs. The substantial difference between this peak and the average accounts for the large standard deviation of 19.44. In contrast, the lowest successful epochs were recorded with the combination of chroma-STFT and pure EffNet, achieving only 11 of the 2000 epochs, with an average and standard deviation of 28.80 and 10.47, respectively. However, Figure 35 suggests that the addition of LSTM-based architectures can effectively delay the premature activation of the early stopping mechanism. In the figure, both EffNet-BiLSTM and SWinT-BiLSTM generally required more epochs, on average, than their vanilla and GRU-enhanced versions, independent of the features used.

Figure 36 presents a comprehensive analysis detailing the average training epochs required by several hybrid DL models on the ESC-50 dataset. This analysis incorporates three distinct methods of feature extraction. The six DL models blend foundational CNNs (EffNet) and transformer-based architectures (SWinT) with RNN layers such as GRU and LSTM. The figure utilizes color-coding to differentiate epochs based on input feature types, while error bars indicate the standard deviation across the 5-fold CV runs, following the common evaluation protocol used with ESC-50 [5]. Notably, the EffNet-BiLSTM model employing MFCCs features exhibited the highest training duration, averaging 556.2 epochs with a standard deviation of 62.79, reflecting its extended convergence time and sensitivity to compact spectral representations. Similarly, when using mel-spectrogram features, EffNet-BiLSTM required an average of 440.6 epochs, underscoring the persistence of deep LSTM-based architectures before early stopping is triggered. In contrast, within the EffNet family, the baseline

EffNet model with the mel-spectrogram converged more rapidly, averaging 188 epochs with a low deviation of 9.67, indicating comparatively fast convergence. Overall, the number of required epochs varies substantially depending on both the model architecture and the feature type. The lowest mean epoch count is observed for the SWinT-BiGRU model with chroma-STFT, averaging 168.8 epochs, demonstrating rapid convergence for this specific model–feature pairing. In conclusion, Figure 36 highlights the strong influence of the architectural complexity and feature richness on the training duration. Models incorporating LSTM layers generally require longer training periods, reflecting their enhanced capacity to model temporal dependencies, albeit at the cost of an increased computational demand. Moreover, due to the balanced nature of the ESC-50 dataset, as indicated in Table 8, training extends over a greater number of epochs compared to imbalanced datasets such as MIMII, even though the overall accuracy remains lower, as illustrated by Figures 16 and 17.

Lastly, Figure 37 provides a comparison of the average training epochs for different hybrid DL models on the FSC22 dataset. The figure evaluates three feature representations: chroma-STFT, MFCC, and the mel-spectrogram, across six model architectures that combine CNN-based (EffNet) and transformer-based (SWinT) backbones with recurrent units such as GRU and LSTM layers. The error bars represent the standard deviation obtained from five-fold cross-validation, following standard evaluation practices [5]. In particular, the EffNet-BiLSTM model using MFCC features exhibits the highest mean epoch count at 348.2, suggesting prolonged optimization before early stopping, likely due to the interaction between compressed spectral features and deep temporal modeling. A similar pattern is observed for EffNet-BiLSTM with the mel-spectrogram, requiring an average of 303.0 epochs, reflecting the increased training demands of LSTM-augmented architectures. In contrast, the SWinT-BiGRU model consistently converges earlier across all feature types, with the fastest convergence observed when paired with chroma-STFT at 126.0 epochs. Additionally, the baseline EffNet model with MFCC converges more quickly than its LSTM-enhanced counterpart but exhibits higher variance, indicating a less stable learning trajectory. Overall, models using chroma-STFT features tend to trigger early stopping sooner than those employing MFCC or the mel-spectrogram. The figure further indicates that LSTM-based models (EffNet-BiLSTM and SWinT-BiLSTM) generally require more epochs before reaching optimal validation performance, which is consistent with their stronger temporal modeling capacity and the improved performance observed for the best-performing configurations in earlier evaluation metrics.

3.3. Discussion

This section discusses and contextualizes the experimental results from comparing different feature representations and model architectures. Although mel-spectrogram-based representations typically deliver the most accurate and consistent performance across both industrial and urban benchmarks, their advantage does not hold uniformly. In acoustically heterogeneous natural environments,

such as those captured in FSC22, their effectiveness becomes strongly dependent on the specific characteristics of the dataset. It is important to note that the term 'generalization' in this context refers to model robustness evaluated within each dataset under consistent cross-validation, rather than cross-dataset or TL scenarios.

3.3.1. Outcomes According to Assessment Criteria

Metric-based evaluation on the MIMII, ESC-50, and FSC22 datasets consistently indicates that strong audio classification performance depends heavily on carefully matching robust spectral representations with architectures that can model temporal dynamics. Among the feature sets tested, mel-spectrograms appeared in nearly all top-scoring configurations, particularly when combined with transformer-based models such as SWinT augmented with recurrent layers (e.g., SWinT-BiLSTM). This effectiveness arises from their detailed time–frequency resolution and the ability of these models to capture long-range temporal dependencies, as shown in Figures 16 to 31. In contrast, features like chroma-STFT, while less computationally demanding, frequently lagged behind due to their lower resolution, especially in noisy or non-tonal settings. Likewise, baseline architectures such as EffNet generally fell short of hybrid models in terms of precision. It is important to emphasize, however, that architectures achieving the highest precision—particularly those combining self-attention with recurrence—also incur greater computational cost and longer training times. These setups demand more powerful hardware, limiting their suitability for resource-constrained or real-time applications unless additional optimization is applied. Consequently, although advanced architectures and expressive feature representations significantly enhance classification performance, they also pose challenges for scalability, latency, and energy efficiency. Balancing accuracy with computational efficiency, therefore, remains central to the practical design and deployment of DL-based environmental audio classification systems. While the number of training epochs alone is not a definitive indicator of overfitting, the convergence patterns observed under a uniform early-stopping scheme offer a useful comparative signal of training stability across different architectures.

3.3.2. Performance Enhancement through Integration of RNN-based Model

The evaluation of all metrics (i.e., precision, AUC, etc.) indicates that the combination of SWinT and mel-spectrogram in the classification of the data set MIMII consistently reaches the best score among other alternatives involved in this study. The addition of layers BiGRU and BiLSTM to the SWinT baseline model does not improve the discrimination quality and even slightly deteriorates the final products. However, the insertion of BiLSTM before the final classification layers of EffNet consistently helps to improve the quality of the achievements in MIMII's classification. The improvement constantly occurs regardless of the feature employed. In addition, the use of MFCC, although it is often placed in the middle position among the three variants, should be considered, particularly for classification with limited

computational resources.

Another closer inspection of Figures 16 and 20 indicates that adding recurrent layers, such as BiGRU and BiLSTM, to baseline models, in this case, EffNet and SWinT, does not necessarily enhance performance, particularly when combined with the mel-spectrogram. Even worse, their introduction to transformer-based variants (i.e., SWinT-BiGRU, SWinT-BiLSTM) caused a deterioration in both mean accuracy and AUC compared to standard SWinT. This implies that the transformers' intrinsic ability to model long-term dependencies might suffice when combined with high-dimensional features, making RNN augmentations unnecessary or potentially detrimental. However, a different development emerges when the BiGRU and BiLSTM layers are fused with EffNet, specifically in conjunction with the use of MFCC or chroma-STFT. In such cases, there were modest performance improvements, suggesting that the sequential modeling strengths of RNNs can provide supplementary advantages when feature representations are either less expressive or more compact. These findings collectively underscore the need for careful calibration between the feature type and the model architecture, as certain combinations can enhance or reduce classification effectiveness depending on the input data richness.

Incorporating recurrent components, such as BiLSTM and BiGRU, into foundational architectures yields varying outcomes depending on the model's complexity and the expressiveness of the features. As shown in Figures 17, 21, 24, 27, and 30, hybrid models such as EffNet-BiLSTM and EffNet-BiGRU generally improve performance in classification ESC-50. For example, integrating MFCC with EffNet and BiLSTM raises accuracy from $55.05\% \pm 5.06\%$ to $68.35\% \pm 4.51\%$ and AUC from $96.33\% \pm 0.78\%$ to $98.72\% \pm 0.27\%$, demonstrating enhanced temporal modeling with less detailed features. For specific cases, the statistical performances (i.e., accuracy, AUC, precision, recall, and F1) of EffNet-BiLSTM even exceed those of SWinT and SWinT-BiGRU, both of which use MFCC features. The combination of EffNet-BiLSTM and MFCC even has the ability to overtake the product of EffNet-BiLSTM and the mel-spectrogram. Meanwhile, in the implementation of chroma-STFT, both EffNet-BiGRU and EffNet-BiLSTM successfully surpass the achievements of the foundational model. These recurrent components continue to offer comparable advantages for more complex features such as the mel-spectrogram. When augmented with BiLSTM, for example, the SWinT-based models also show performance improvements. The standard SWinT achieves $84.00\% \pm 2.14\%$ accuracy and $99.55\% \pm 0.08\%$ AUC with the mel-spectrogram, while SWinT-BiLSTM increases accuracy to $91.15\% \pm 1.67\%$ and AUC to $99.93\% \pm 0.03\%$. Unfortunately, these advantages are not always made clear in all modifications. In the case of MFCC, EffNet-BiGRU slightly underperforms the plain EffNet, indicating that recurrent layers may add redundancy or alter sequence dependency modeling compared to attention mechanisms. This suggests that recurrent layers are most effective in scenarios with less complex features, as their utility declines in models with advanced temporal encoding, such as transformers. Thus, building hybrid architectures should

strike a balance between the feature detail and model capacity, avoiding unnecessary complexity that increases computational demand without notable performance gains. These discoveries are in line with Andayani et al. (2022) [183]'s findings, who also observed significant performance enhancements when combining the BiLSTM and Transformer layers in hybrid models, particularly with less temporally detailed input features, while enhancements diminish with an increasing expressiveness of the feature.

Incorporating RNN layers, such as GRUs and LSTMs, into foundational models results in substantial modifications to their performance metrics and helps delay overfitting. The magnitude of these effects is heavily influenced by the model's core architecture and the specific nature of the feature representations employed. For simpler feature representations, like chroma-STFT and MFCC, the addition of temporal modeling layers significantly elevates the classification accuracy and enhances the AUC. For instance, the application of EffNet-BiLSTM results in a remarkable increase in accuracy, advancing from the baseline $50.86\% \pm 16.66\%$ observed with the original EffNet using MFCC, reaching an enhanced accuracy of $75.60\% \pm 1.90\%$. Concurrently, the AUC experiences an improvement from $92.59\% \pm 6.54\%$ to a notable $98.95\% \pm 0.25\%$. These outcomes underscore the capacity of recurrent layers to address the limitations in temporal resolution associated with simpler feature sets. Conversely, when using the mel-spectrograms, which inherently offer a comprehensive matrix of temporal and frequency information, the supplementary advantages posed by incorporating GRU or LSTM layers, although still significant, are comparatively subtle. For example, integration of SWinT with the mel-spectrogram results in an accuracy of $75.75\% \pm 2.76\%$ and an AUC of $98.57\% \pm 0.26\%$. By extending this setup to include BiLSTM within SWinT-BiLSTM, the performance metrics are further elevated, achieving an accuracy of $83.16\% \pm 0.47\%$ and an AUC of $99.55\% \pm 0.04\%$. In contrast, for features like chroma-STFT, the gains from embedding BiGRU or BiLSTM are relatively constrained. This constraint is potentially attributed to the feature's inadequacy in effectively capturing spectral information, which may fail to be fully compensated for by temporal modeling. The evidence suggests that while the inclusion of recurrent layers indeed augments model performance, their benefits are most effectively leveraged in models dealing with low-to-intermediate-level feature sets. For highly detailed features such as mel-spectrograms, the supplementary benefits must be weighed cautiously against the added complexity they contribute to the model. The evaluation of these performance metrics was conducted utilizing a standardized k-folds CV method along with an early stopping criterion, ensuring a thorough assessment framework and safeguarding against overfitting during model training.

3.3.3. Impact of Employing Different Feature Extraction Methods

An examination of the classification trends presented in Figures 16 and 20 highlights the key impact of feature representation on the model performance across

various DL frameworks. In particular, integration of the mel-spectrogram with models based on SWinT consistently produces superior results, supporting the notion that detailed time-frequency features work satisfactorily with attention-based architectures. For example, although chroma-STFT is relevant in musical contexts, it achieves a maximum mean score of just 81.41% when paired with SWinT, which is nevertheless lower than the minimum accuracy obtained with MFCC or the mel-spectrogram. This supports the idea of its lesser suitability in intricate industrial audio contexts. This underperformance is related to the dimensional constraints of chroma-STFT, as described in Table 11, which produces significantly fewer coefficients than the mel-spectrogram, which is more than 10 times fewer in terms of the dimension's depth, thus limiting its ability to capture intricate temporal and spectral features of machine-generated sound patterns.

Meanwhile, the analysis of feature extraction techniques in Figures 17, 21, 24, 27, and 30 clearly illustrates how the selected audio representation substantially influences the effectiveness of models to classify environmental sounds. Of the three assessed characteristics, the mel-spectrogram consistently surpasses both MFCC and chroma-STFT in accuracy and AUC, achieving top performance of $91.15\% \pm 1.67\%$ and $99.93\% \pm 0.03\%$, respectively, when used in conjunction with SWinT-BiLSTM. This underscores the ability of the mel-spectrogram to capture intricate time-frequency structures crucial for identifying complex sound categories. Although the MFCC features secure a middle ground, benefiting from their compact, perception-focused design, SWinT-BiLSTM still delivers optimal results for MFCC-based models, with $72.30\% \pm 2.41\%$ recall and an AUC of $99.03\% \pm 0.31\%$. However, MFCC might miss some of the finer spectral nuances preserved in the mel-spectrogram despite their computational efficiency. Chroma-STFT, on the other hand, which focuses on harmonic and pitch-class information, is the least effective in this scenario, as evidenced by its limited ability to capture a wide variety of non-musical environmental sounds, with only $54.70\% \pm 2.55\%$ accuracy even in the most sophisticated model setup. Consequently, these findings indicate that the mel-spectrogram is more preferable in general environmental audio classification contexts such as ESC-50, while MFCC strikes a balance between performance and computational efficiency, and chroma-STFT might be better-suited to tonal or music-focused applications, rather than comprehensive environmental soundscapes. Nevertheless, this discovery aligns with Lorena et al.'s (2020) study [184] that as the number of classification categories increases, precision and computational demands rise.

In case of FSC22 classification, the assessment of audio features, namely, chroma-STFT, MFCC, and the mel-spectrogram, reveals significant disparities in their efficacy to classify environmental sounds from the wilderness, such as those present in FSC22. In particular, chroma-STFT, which emphasizes pitch classes and harmonic content [157, 185, 186], often yields the lowest scores among all metrics in various models. This information suggests that chroma-STFT is deficient in the necessary spectral and temporal detail required to effectively discriminate the diverse

range of natural and anthropogenic sounds encountered in forest environments, although it is well tailored for musical applications. In contrast, MFCC concentrates on more perceptually relevant information, leading to moderate improvements in classification results. For example, in conjunction with EffNet-BiLSTM, MFCC achieves a mean accuracy of $75.60\% \pm 1.90\%$, surpassing chroma-STFT and other integrations of model features. However, the most pronounced results are achieved with mel-spectrograms, which systematically yield the highest scores for all metrics in almost every configuration. Their advanced time-frequency representation allows the model to discern subtle variations in the sound texture and duration, which are crucial for identifying intricate forest acoustics. These findings underscore the crucial importance of selecting feature representations that align with the acoustic variability of the environment under study. In this context, the exceptional performance of the mel-spectrogram in all metrics confirms its suitability for bioacoustic analysis and environmental sound classification.

Beyond the empirical trends reported above, the contrasting behavior of MFCC and the mel-spectrogram highlights a broader methodological shift in how acoustic front-ends interact with the inductive biases of modern model architectures. Traditionally, MFCCs were preferred because the cepstral DCT produces a compact, approximately decorrelated feature space that aligns well with classical generative models such as the Gaussian mixture model (GMM)–hidden Markov model (HMM), which typically rely on diagonal covariance assumptions for computational tractability [187, 188]. In contemporary DL systems, however, strict decorrelation is no longer a prerequisite: convolutional and attention-based networks can directly leverage a correlated time–frequency structure and are able to exploit richer local patterns (e.g., spectro-temporal edges, onset cues, fine-grained band-energy fluctuations) that are often suppressed by cepstral compression. This perspective helps clarify why mel-spectrograms systematically outperform MFCCs in Figures 16 and 17, and why this performance gap is echoed in the more distinct cluster separations observed in the t-SNE embeddings (e.g., Figures 33c and 34c). The broader body of work is consistent with this trend: recent high-performing transformer-based and self-supervised audio models are predominantly trained on log-mel or mel-filterbank spectrograms, leveraging their higher information content and compatibility with patch-based or convolutional input stages [86, 88, 189–191]. Consequently, the superior performance of mel-spectrograms observed here is unlikely to be an artifact of the specific datasets; rather, it indicates that richer time–frequency representations are generally better aligned with modern neural architectures, whereas MFCC remains a compelling low-dimensional choice primarily when computational budgets or limited model capacity rule out high-resolution spectrogram inputs (see Table 11). Taken together, these results suggest that MFCC should not be regarded as an inherently weaker representation, but as a computationally efficient and robust alternative that is particularly appealing when the model size, available training data, or deployment resources are tightly constrained.

3.3.4. Epoch and Convergence Analysis

The comparative epoch analyses in Figures 35, 36, and 37 highlight a key trade-off between rapid convergence and the danger of stopping training too early. Models that require fewer epochs reach their minimum loss more quickly, indicating efficient learning; however, if training halts before the model has fully captured the discriminative structure, this can also signal underfitting. This behavior is clear in models such as EffNet on MIMII and SWinT-BiGRU on ESC-50 and FSC22, which trigger early stopping after relatively few epochs. For instance, the SWinT-BiGRU model using chroma-STFT features on both MIMII and FSC22 converged sooner than other configurations, suggesting that its architecture is well-suited to simpler or lower-dimensional representations. At the same time, this early convergence may also indicate a restricted capacity to learn more complex relationships, potentially weakening generalization on difficult or highly variable data, particularly in imbalanced datasets such as MIMII. By contrast, models incorporating BiLSTM layers (e.g., EffNet-BiLSTM and SWinT-BiLSTM) tended to train for more epochs, especially with MFCC or the mel-spectrogram inputs, reflecting their ability to capture richer temporal dependencies. This prolonged training can support more thorough optimization but demands greater computational cost and tighter regularization. Overall, while fewer required epochs point to faster convergence, this can either represent effective early generalization or an early stopping point that ultimately constrains the model's capacity to generalize robustly to unseen data.

3.3.5. Comparison to Previous Studies

Tables 33, 34, and 35 provide a comparative overview of the highest reported performance levels in recent studies addressing classification tasks on the MIMII, ESC-50, and FSC22 datasets. These summaries establish an essential context for positioning the results of this study relative to existing state-of-the-art approaches.

Table 33 shows that a large fraction of existing studies assess performance under relatively narrow experimental conditions (e.g., targeting specific machine types [192, 194, 196], limiting the task to binary normal–abnormal classification [195], or constraining training and validation to fixed SNR levels [23, 46]). While these methods report competitive accuracies between 86.44% and 94.49%, their restricted setups hinder direct comparison. In contrast, the proposed method attains an accuracy of 99.06% by leveraging SWinT with mel-spectrograms over the entire MIMII dataset and adopting a unified k-folds CV strategy for training and validation. This finding aligns with the earlier analyses in Section 3.1 and Subsection 3.2.1, reinforcing that pairing rich time–frequency representations with transformer-based models delivers strong performance on MIMII. At the same time, the results indicate that adding recurrent modules such as BiLSTM on top of SWinT offers only marginal gains. The use of early stopping together with cross-validation further enhances robustness and helps prevent overfitting, which is consistent with the observations reported in [183].

Table 33. Comparison of the highest achievements of selected recent studies of MIMII classification

Author(s)	Method	Highest Accuracy	Notes
Ding et al. (2023) [192]	CNN-based model	94.25% (slider)	Conducting classification on operation states of each machine separately
Siraj et al. (2023) [193]	Few-shot learning, MobileNet, and STFT	86.44%	Focus on abnormal collections and implementation of the model on their self-collected data
Pu et al. (2023) [194]	IEMD-DDCNN	94.49%	Focus only on pump data without ID distinction
Alagele et al. (2024) [195]	AE and MFCC	93.95%	Only divided into normal and abnormal classes
Chandrakala et al. (2024) [196]	Spectro Temporal Fusion with CLSTM-Autoencoder	92.94% (Slider)	Conducting classification on operation states of each machine separately
Zabin et al. (2024) [23]	Self-Attention SqueezeNet	89.32%	Training and validation only data with -6 dB SNR
Zabin et al. (2025) [46]	Few-shot learning with EMD-Gammatone Spectrogram	89.6%	Training and validation only data with -6 dB SNR
This study	EffNet, SWinT, and proposed models	99.06%	Training and validation with the whole data in the dataset

Table 34. Comparison of the highest achievements of selected recent studies of ESC-50 classification

Author(s)	Method	Highest Accuracy
Lin et al. (2020) [137]	ParallelNet	81.55%
Zhou and Zhao (2022) [138]	TIANnet	84.2%
Li et al. (2022) [139]	Attention-based CNN with MFR feature	93.1%
Gong et al. (2023) [140]	Audio Classification Method based on Self-Supervised and Knowledge Distillation	97.2%
Liu et al. (2023) [91]	CAT with PANN	96.9%
Sarkar and Etemad (2022) [197]	CrissCross	79%
Chen et al. (2024) [198]	contrastive learning-based audio spectrogram transformer (CL-Transformer)	97.75%
Ranmal et al. (2024) [143]	ESC-NAS	81%
Chen et al. (2025) [69]	MobileNetV2 + SPA	91.75%
This Study	EffNet, SWinT, and proposed models	93.50%

Meanwhile, the top reported accuracies on the ESC-50 benchmark surpass 97%, with Chen et al. [198] attaining 97.75% through CL-Transformer, as summarized in Table 34. These results are generally achieved by leveraging extensive pretraining, self-supervised learning, or knowledge distillation techniques [91, 140, 198]. By contrast, the present study obtains an accuracy of 93.50% using SWinT-BiLSTM in combination with mel-spectrograms. While this accuracy remains below the current state-of-the-art, it is produced within a consistent and unified experimental setup applied across multiple datasets, enhancing the robustness and fairness of the comparison. The results further support previous findings that mel-spectrograms outperform MFCC and chroma-STFT representations on ESC-50. Moreover, integration of a BiLSTM component into the SWinT architecture provides clear performance improvements (Figures 17–30). The residual performance gap with respect to leading methods largely reflects methodological design decisions, as this work intentionally forgoes large-scale pretraining to maintain controlled model complexity and experimental uniformity [184].

Table 35. Comparison of the highest achievements of selected recent studies of FSC22 classification

Author(s)	Method	Highest Accuracy
Bandara et al. (2023) [7]	CNN-based model with the mel-spectrogram	92.59%
Ahmad et al. (2024) [146]	Stacking method with MFCC	72.7%
Ranmal et al. (2024) [143]	ESC-NAS	85.78%
Simiyu et al. (2024) [199]	Temporal Frequency CNN	87%
Xu and Chen (2024) [144]	ERT	66%
Qurthobi et al. (2025) [148]	CNN-BiLSTM with MFCC	78.52%
Sims et al. (2025) [200]	ZeroDiffusion	39.75%
This Study	EffNet, SWinT, and proposed models	83.16%

For the FSC22 dataset, Table 35 shows that reported accuracies span from 66% to 92.59%. Within this range, the proposed method achieves an accuracy of 83.16% using SWinT-BiLSTM with mel-spectrograms, thus surpassing several recent baselines [146, 148, 199, 200]. However, it still falls short of the highest reported performance by Bandara et al. [7]. These discrepancies likely arise from differing evaluation protocols, class definitions, and augmentation strategies. In line with the convergence analysis in Subsection 3.2.1, mel-spectrograms again prove to be the most effective feature representation, whereas MFCC yields intermediate results and chroma-STFT consistently performs poorly, reflecting its limited suitability for non-tonal acoustic settings [157, 185, 186]. The performance gains obtained by adding recurrent layers to SWinT further underscore the critical role of temporal modeling in forest bioacoustic classification, albeit with a higher computational cost.

Although these patterns are most evident on the MIMII and ESC-50 benchmarks, FSC22 exhibits greater variability owing to its richer acoustic diversity and temporal

complexity. First, the choice of input representation is pivotal: mel-spectrograms combined with attention-based architectures consistently achieve the strongest performance, underscoring the benefit of coupling detailed time–frequency features with self-attention mechanisms. Second, incorporating recurrent layers yields mixed outcomes; they enhance results for compact representations such as MFCC and chroma-STFT, yet provide only modest improvements for transformer-based models that already capture long-range dependencies [183]. These gains are thus highly context-sensitive, most notable for convolutional backbones and compact feature sets, but largely marginal for transformer architectures that inherently encode temporal structure. Third, evaluation methodology critically influences reported performance. Whereas much prior work adopts device-specific or SNR-restricted configurations, this study applies full k-folds CV across entire datasets, thereby improving both fairness and reproducibility. Finally, even though state-of-the-art ESC-50 accuracies exceed 90%, they often rely on extensive pretraining and substantial computational budgets, highlighting the tension between the peak accuracy and practical deployability in resource-limited scenarios.

3.3.6. Computational Loads

Beyond conventional performance metrics, the practicality of an audio classification model also hinges on its computational costs. For deployments on resource-constrained devices such as embedded sensors, smartphones, or industrial edge platforms, training and inference efficiency are as important as predictive accuracy [86, 183]. Factors like per-epoch training time, model size, and hardware demands directly affect scalability, latency, and power consumption, and thus determine suitability for real-world use [24, 91]. This work examines the computational requirements of the proposed models across multiple datasets and feature types, complementing the prior accuracy-focused evaluation.

Table 36. Computational times during MIMII classification

Model	Computation time per epoch (seconds)		
	mel-spectrogram	MFCC	chroma-STFT
EffNet	~ 80	~ 40	~ 40
EffNet-BiGRU	~ 90	~ 50	~ 50
EffNet-BiLSTM	~ 95	~ 55	~ 55
SWinT	~ 120	~ 50	~ 45
SWinT-BiGRU	~ 130	~ 55	~ 50
SWinT-BiLSTM	~ 130	~ 60	~ 55

The experiments discussed above were primarily designed to assess classification accuracy; however, for practical deployment, computational overhead and resource usage are equally crucial considerations. Tables 36–38 summarize the per-epoch training durations for each dataset–feature combination, which we use as an indicator of inference complexity. Transformer-based architectures (i.e.,

Table 37. Computational times during ESC-50 classification

Model	Computation time per epoch (seconds)		
	mel-spectrogram	MFCC	chroma-STFT
EffNet	~ 3	~ 2	~ 2
EffNet-BiGRU	~ 4	~ 2	~ 2
EffNet-BiLSTM	~ 4	~ 2	~ 2
SWinT	~ 4	~ 2	~ 1
SWinT-BiGRU	~ 8	~ 3	~ 2
SWinT-BiLSTM	~ 8	~ 3	~ 3

Table 38. Computational times during FSC22 classification

Model	Feature		
	mel-spectrogram	MFCC	chroma-STFT
EffNet	~ 3	~ 3	~ 1
EffNet-BiGRU	~ 4	~ 2	~ 2
EffNet-BiLSTM	~ 4	~ 3	~ 2
SWinT	~ 4	~ 2	~ 1
SWinT-BiGRU	~ 8	~ 3	~ 3
SWinT-BiLSTM	~ 8	~ 3	~ 3

SWinT, SWinT-BiGRU, SWinT-BiLSTM) require more computation time than their CNN-based counterparts. This indicates that, while attention-driven models typically yield superior classification performance, they do so at the cost of increased computational requirements [24, 91].

The computational time depends heavily on the representation of the features. The mel-spectrograms, with their high time–frequency resolution, require the longest training durations (up to $\sim$ 130 seconds per epoch on MIMII), while MFCC and chroma-STFT cut these times nearly in half. On smaller datasets (ESC-50, FSC22), the absolute times reduce to a few seconds per epoch, but the same relative ordering holds. This pattern highlights the trade-off: richer features like mel-spectrograms yield superior accuracy (see Tables 33–35) but increase the computational burden [157, 184].

These measurements were obtained on a workstation with the hardware and software setup described in Section 2.6, and therefore only approximate the true computational cost. While the models can run on such desktop-class machines at acceptable expense, their resource demands are likely unsuitable for edge devices, where compute, memory, and energy are severely limited. As a result, despite the proposed hybrids, including SWinT-BiLSTM, achieving SOTA accuracy on the evaluated datasets, their inference cost remains a major obstacle for embedded or battery-powered deployments, particularly under continuous or real-time constraints. Addressing this issue will require future work on approaches such as model compression, low-precision quantization, and structured or unstructured pruning [7,

183], possibly in combination, to maintain predictive quality while meeting stringent latency, memory, and power budgets typical of real-time and low-resource scenarios.

3.3.7. Limitations of Study

Although this study achieved commendable accuracy on MIMII, ESC-50, and FSC22, it has limitations, particularly in practical use. The suggested hybrid models (EffNet-BiGRU, EffNet-BiLSTM, SWinT-BiGRU, SWinT-BiLSTM) outperformed baselines in tested datasets, but their generalization to unfamiliar acoustic domains remains doubtful. Despite the datasets' diversity, they focus on specific industrial, urban, and natural environments, potentially excluding the full complexity of real-world audio settings. For instance, the models might struggle in highly dynamic or unpredictable noise conditions, such as sudden machine breakdowns, extreme weather events, or overlapping human and natural sounds, as they were trained on more orderly data.

High-resolution spectral features, such as mel-spectrograms, present additional challenges for transferability. Although these features deliver precise time-frequency details and perform well in tested domains, they might also capture dataset-specific noise patterns, risking overfitting. This raises concerns about model performance stability when faced with new recording conditions, new sensors, or different microphone arrangements. Prior research has indicated that acoustic classification systems may experience notable performance drops under mismatched domain conditions [38, 91].

Computationally, the study revealed that the models with the greatest accuracy (SWinT-BiLSTM using mel-spectrograms) also have the highest processing demands (see Tables 36–38). Although these models operate well on high-performance GPUs, they may not be ideal for low-power or edge devices unless optimized. Methods such as quantization, pruning, or knowledge distillation could mitigate these issues [7, 183]; however, they were not investigated in this study. Therefore, future work should evaluate both domain generalization and computational efficiency to move from controlled environments to reliable, real-world applications successfully.

3.3.8. Classifications with Hybrid Approaches

This subsection examines initial evidence on the effectiveness of hybrid feature representations by systematically comparing them with individual feature extractors, namely, mel-spectrogram, MFCC, and chroma-STFT. The analysis focuses on the maximum performance values reported in the tables and figures, as these reflect the best achievable outcomes under identical training and evaluation conditions. By contrasting the strongest single-feature models with the best hybrid configurations, we assess whether feature combinations provide benefits beyond those attributable to the network architecture. This issue is particularly pertinent for anomaly event detection, where improvements may be small but practically meaningful, especially in noisy settings with rare and weakly expressed abnormal events. Similar hybrid strategies

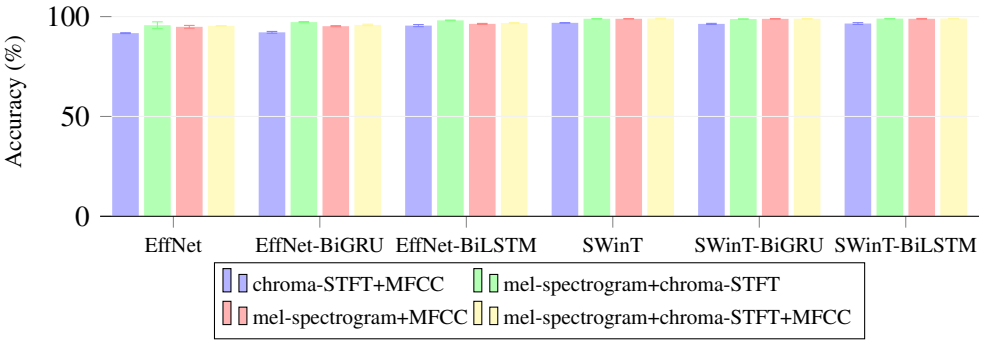
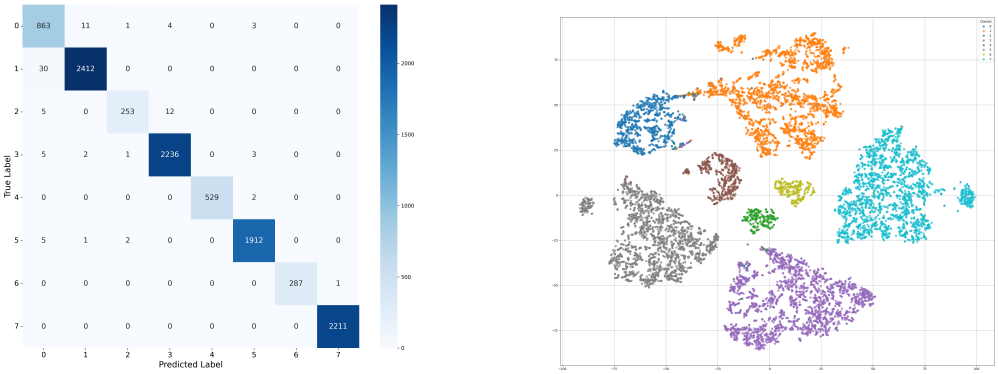


Fig. 38. Comparative statistical analysis of classification accuracies on the MIMII dataset using hybrid extractors



(a) Confusion matrix

(b) t-SNE

Fig. 39. Confusion matrix and t-SNE visualization during the highest achievement of MIMII classification with hybrid extractors and SWinT-BiLSTM

have previously been shown to enhance robustness and discriminative capability in demanding acoustic tasks, such as ESC and industrial condition monitoring [4, 5, 7].

In the industrial acoustic scenario, a direct comparison of Figure 38 with Figure 16 shows that both the baseline feature sets and the hybrid feature combinations yield near-ceiling performance on the MIMII dataset. The strongest individual feature configuration achieves an average accuracy of $99.02\% \pm 0.02\%$. When multiple feature extractors are fused into a single representation and used together with SWinT-BiLSTM, the best hybrid setup offers only a marginal improvement, reaching $99.03\% \pm 0.10\%$. The highest accuracy observed for any single experiment, notably, 99.18%, corresponds to the confusion matrix and the t-SNE embedding visualized in Figure 39.

Consistent with the observations for single-extractor scenarios, attention-based

architectures provide superior classification performance compared to purely CNN-based models. Although the absolute gain of about 0.01 percentage points is practically negligible, it must be interpreted in the context of an already highly optimized classifier, where ceiling effects naturally constrain further measurable improvements. Importantly, the hybrid approach maintains this top-level accuracy while exploiting a richer, more diverse feature representation, demonstrating that feature fusion does not degrade anomaly detection capability, even when performance is already close to the upper bound.

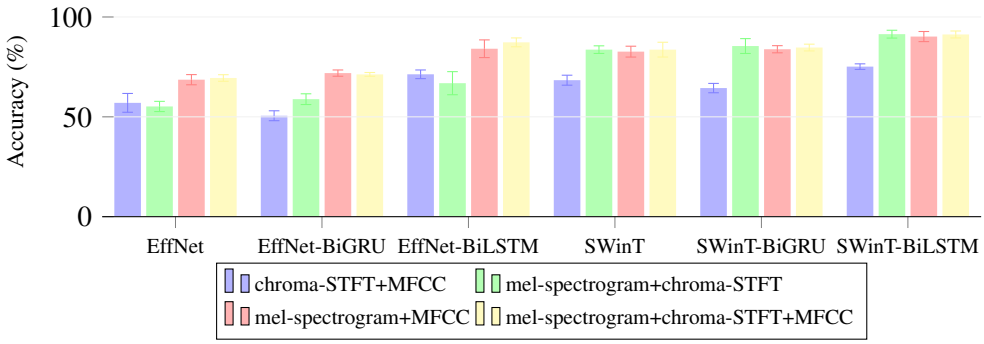
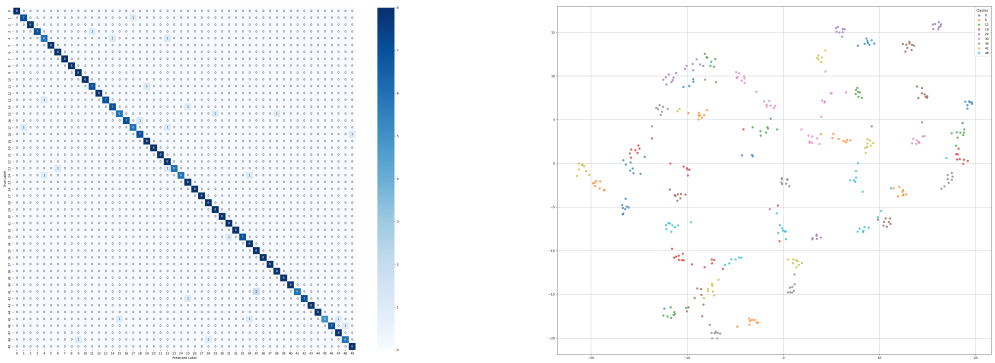


Fig. 40. Comparative statistical analysis of classification accuracies on the ESC-50 dataset using hybrid extractors



(a) Confusion matrix **(b)** t-SNE

Fig. 41. Confusion matrix and t-SNE visualization during the highest achievement of ESC-50 classification with hybrid extractors and SWinT-BiLSTM

From the perspective of anomaly detection, these findings indicate that hybrid feature representations remain effective at identifying subtle irregularities in machine operation—for instance, minor shifts in vibration signatures or changes in mechanical resonance—while at the same time offering increased robustness to fluctuations

in ambient acoustic environments and background operational noise. Comparable tendencies have been documented in the wider field of industrial acoustic condition monitoring, where research and practice generally prioritize long-term stability, reliability, and robustness of detection systems over incremental gains in conventional accuracy metrics [4, 23]. This broader body of work reinforces the idea that, in real-world industrial deployments, the ability to consistently detect anomalies under varying environmental and operational conditions is often valued more highly than small improvements in benchmark performance.

In contrast, the urban and suburban soundscapes, represented by ESC-50, gain substantially more from the introduction of hybrid features in quantitative terms. As can be observed by comparing Figure 40 with Figure 17, the strongest setups that rely exclusively on unprocessed, pristine features reach maximum accuracies only in the high-80% to low-90% interval. Once hybrid features are incorporated, however, the top-performing systems consistently shift into the low- to mid-90% accuracy range.

More specifically, the configuration that combines all available feature extractors with SWinT-BiLSTM attains the best overall results, achieving an average accuracy of $91.25\% \pm 1.74\%$ and a peak single-run accuracy of 93.25%. This leading setup is visualized in Figure 41, which presents its confusion matrix and the associated t-SNE embedding. Taken together, these findings indicate an absolute gain of approximately 2–3 percentage points at the upper end of the achievable performance range.

For a dataset characterized by numerous acoustically similar categories and a wide variety of sound sources, such gains are far from marginal. The elevated peak accuracy indicates that hybrid representations enable the model to jointly exploit complementary spectral, cepstral, and harmonic cues, leading to finer discrimination of infrequent, ambiguous, or atypical acoustic patterns. From the perspective of anomaly analysis, this is particularly valuable, because anomalous events in urban environments generally appear as rare, contextually incongruous sounds that are partially masked by persistent background noise—examples include unusual human behaviors, unexpected mechanical failures, or sporadic environmental incidents. In line with our observations, prior work on urban sound recognition also reports that multi-feature fusion mitigates inter-class confusion and strengthens robustness to overlapping sources and noise [5, 6, 201].

A similar, albeit smaller, improvement is also present under natural acoustic conditions. When comparing Figure 42 with Figure 19, it becomes evident that the best individual feature set generally attains a maximum accuracy in the range of approximately 70%–80%. By contrast, the most effective hybrid configuration consistently elevates this maximum by around 1–2 percentage points, indicating a modest yet systematic performance benefit.

When FSC22 is trained by using a hybrid feature extractor that integrates mel-spectrogram, MFCC, and chroma-STFT, the advantages of EffNet-BiLSTM in terms of efficiency and inference speed become even more pronounced. Under these conditions, EffNet-BiLSTM achieves the best performance, obtaining a mean

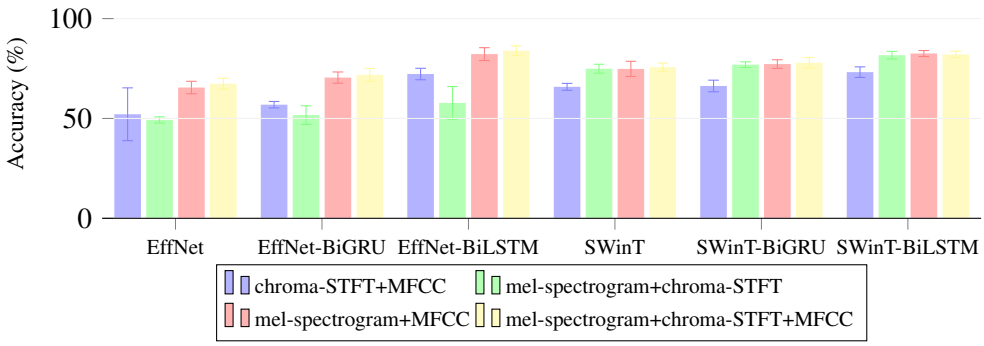
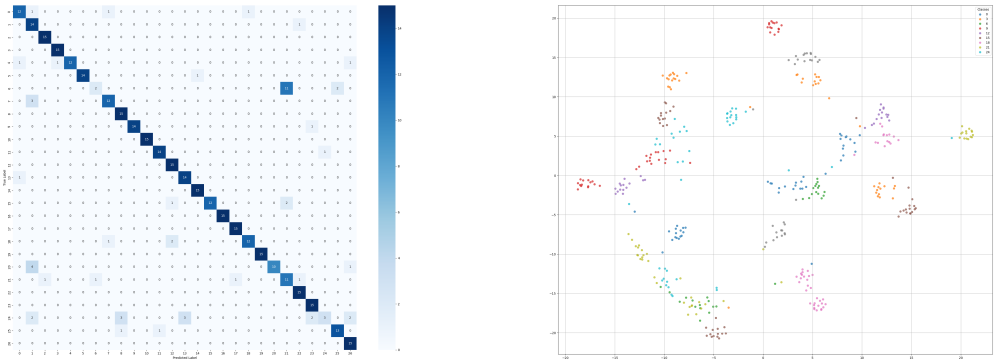


Fig. 42. Comparative statistical analysis of classification accuracies on the FSC22 dataset using hybrid extractors



(a) Confusion matrix **(b)** t-SNE

Fig. 43. Confusion matrix and t-SNE visualization during the highest achievement of FSC22 classification with hybrid extractors and EffNet-BiLSTM

accuracy of $83.95\% \pm 2.43\%$ and outperforming all competing architectures. The confusion matrix and t-SNE projection shown in Figure 43 further confirm that, in contrast to the MIMII and ESC-50 datasets where attention-based models typically dominate, EffNet-BiLSTM emerges as the top-performing approach on FSC22.

Although the absolute accuracies for FSC22 remain lower than those obtained on industrial or urban datasets, the improvement is still notable when the inherent complexity of forest soundscapes is taken into account. Natural ecosystems are frequently characterized by continuous, often dense background noises (e.g., wind, insects, distant animal activity), which are not common features in industrial environments or urban development settings.

The observed gain in peak accuracy, therefore, suggests that hybrid feature representations are more effective at isolating and modeling these weak, transient signals that would otherwise be masked. In other words, the combination of

complementary feature types appears to enhance the sensitivity to subtle deviations from the acoustic background, improving the detection and classification of rare events. This conclusion is consistent with contemporary work in ecological acoustic monitoring, which stresses the importance of rich, multi-faceted feature spaces for the reliable identification of infrequent events in acoustically complex natural habitats [7, 30].

CONCLUSIONS AND FUTURE WORKS

Conclusions

1. The primary goal of this study was to develop a reliable system for recognizing and classifying audio within diverse environmental settings, and the findings verify that this goal has been met. By comparing feature representations (i.e., MFCC, chroma-STFT, and the mel-spectrogram) alongside advanced DL backbones such as EffNet, SWinT, and their temporal variants, the research demonstrated notable performance benefits in noisy and complex auditory scenarios. For example, in the industrial context of the MIMII dataset, the proposed hybrid methods achieved classification accuracies exceeding 99.0%, surpassing previous benchmarks that reported approximately 89.6% [4]. Correspondingly, on the ESC-50 dataset, models integrating SWinT with temporal layers attained accuracies greater than $72.3\% \pm 2.4\%$, exceeding traditional attention-based transformers by over 10 percentage points [5]. In the wilderness setting, combinations of the mel-spectrogram and SWinT-BiLSTM on the FSC22 dataset achieved accuracies up to $82.5\% \pm 1.0\%$, highlighting the framework's capability to handle overlapping natural sounds [7]. These outcomes confirm that the study effectively addresses domain variability and background noise challenges, providing a dependable and precise framework for audio classification in the evaluated environmental contexts. Although the experimental evaluation is conducted by using supervised classification protocols, the reported results directly support anomaly event detection by enabling reliable discrimination between normal and abnormal acoustic patterns in noisy environments.
2. The MIMII dataset, unlike ESC-50 and FSC22, exhibits substantial imbalance between normal and anomalous samples, not only across machine types but also among identifiers of the same device. This imbalance challenges reliable anomaly detection within the dataset. To address this issue, the study employed an SNR-based augmentation strategy using existing recordings at multiple noise levels (i.e., -6 dB, 0 dB, and 6 dB) to simulate realistic operating conditions. This approach expanded the effective training set and introduced meaningful acoustic variability, resulting in an improved classification performance, with models achieving up to $99.06\% \pm 0.02\%$ accuracy, exceeding prior reported results of approximately 89.6% [23, 46]. Performance gains were observed across mel-spectrogram, MFCC, and chroma-STFT features, indicating the robustness of the augmentation strategy. Notably, improvements were most evident under noisy SNR conditions, where overfitting had previously been observed. These findings demonstrate that SNR-based augmentation is an effective mechanism for mitigating dataset imbalance and improving model reliability in industrial AAD.
3. A comparative analysis of three key audio features (i.e., mel-spectrogram, MFCC, and chroma-STFT) reveals distinct performance trends shaped by their interaction with DL architectures and dataset characteristics. Mel-spectrogram, offering

detailed time–frequency resolution, consistently achieves strong classification performance across the evaluated models, particularly on the FSC22 dataset when combined with SWinT-BiLSTM, where an F1-score of $82.47\% \pm 0.98\%$ was obtained. In the MIMII dataset, mel-spectrograms also enhance fault signal detection when paired with hybrid architectures. MFCC features demonstrate a favorable balance between computational efficiency and accuracy, exhibiting stable performance and reduced variance on ESC-50, where SWinT-BiGRU with MFCC achieved recall and precision above 75%. In contrast, chroma-STFT, which emphasizes tonal information, performs poorly in the evaluated non-tonal audio classification tasks, particularly in the MIMII and FSC22 datasets. For instance, EffNet with chroma-STFT achieved only 35.06% accuracy in one configuration and showed limited consistency. These observations are further supported by t-SNE visualizations, which reveal weaker clustering for chroma-STFT compared to the clearer separability observed for the mel-spectrogram and MFCC. Overall, the results emphasize the importance of aligning feature selection with dataset properties and model capacity.

4. To enhance classification performance, this study applies TL to pre-trained EffNet and SWinT models, followed by architectural extensions incorporating temporal modules such as BiGRU and BiLSTM. The resulting hybrid models (EffNet-BiGRU, EffNet-BiLSTM, SWinT-BiGRU, SWinT-BiLSTM) effectively capture temporal dynamics in audio data, particularly for datasets containing rapidly varying or overlapping sound sources. Notable improvements are observed on the ESC-50 dataset, which contains diverse residential sounds, and on the FSC22 dataset with complex wildlife acoustics. For example, SWinT-BiLSTM with MFCC achieved an F1-score of $72.30\% \pm 2.41\%$ on ESC-50, improving upon the standalone SWinT. Similarly, SWinT-BiLSTM with the mel-spectrogram reached an F1-score of $82.47\% \pm 0.98\%$ on FSC22. These findings indicate that temporal modeling can provide meaningful benefits, particularly when combined with suitable backbone architectures and feature representations.
5. The proposed audio classification framework was evaluated on three datasets (i.e., MIMII, ESC-50, and FSC22), each representing a distinct acoustic scenario: industrial, urban, and wilderness environments. These datasets enable a comparative, within-dataset assessment of the framework under varying noise levels and sound characteristics. Performance was measured by using accuracy, AUC, precision, recall, and the F1-score. Hybrid architectures, particularly SWinT-BiLSTM and SWinT-BiGRU, consistently demonstrated strong performance within individual datasets. For instance, on ESC-50, SWinT-BiLSTM with the mel-spectrogram achieved a recall of 91.15%, while on FSC22, SWinT-BiLSTM reached an F1-score of $82.47\% \pm 0.98\%$. In the MIMII dataset, SWinT-BiLSTM with MFCC attained an F1-score of $94.93\% \pm 0.30\%$. t-SNE visualizations further indicate improved class separability for MFCC and mel-spectrogram, supporting their effectiveness in the evaluated classification

tasks.

6. The experiments also provide insight into training dynamics, particularly the role of early stopping based on validation loss stagnation. In all experiments, training was terminated when validation loss failed to improve over 10 consecutive epochs, limiting overfitting. As shown in Figures 35, 36, and 37, none of the evaluated feature–model combinations reached the predefined maximum epoch count. The highest observed mean epoch values were 66.40 ± 19.44 for the mel-spectrogram with EffNet-BiLSTM, 556.20 ± 62.79 for MFCC with BiLSTM, and 303.60 ± 24.48 for the MIMII, ESC-50, and FSC22 datasets, respectively. While the epoch count alone does not directly reflect the model quality, these observations suggest that recurrent layers can improve training stability and convergence behavior, particularly for convolutional backbones paired with compact features.

Future Works

1. Future studies should broaden DL in audio classification to encompass non-traditional settings, such as underwater environments, where sound propagation significantly differs from terrestrial conditions due to aquatic physical properties. These environments feature frequencies outside the human hearing range, such as marine mammal vocalizations, sonar signals, and noise from ships, geological events, or hydrodynamics. DL applied to underwater acoustics promises advances in marine species classification, vessel identification, and detection of oceanographic anomalies. Given the challenges of underwater acoustic data, future research should take advantage of specialized spectral representations and optimized architectures for temporal dependencies. Hybrid models with attention mechanisms and recurrent layers can effectively discern acoustic patterns in complex underwater soundscapes. This research could lead to intelligent, real-time marine monitoring solutions that aid ecology, environmental protection, and underwater security.
2. Another future focus that should be considered is the development of geographically specialized audio datasets to capture the unique sounds of specific countries, regions, or states. The environmental and human-made sounds differ significantly due to variations in climate, biodiversity, culture, infrastructure, and language. For instance, transcripts from the Baltic's flat terrain are more distinct than those from a mountainous Central European region or a lively Mediterranean city. Such datasets would help adapt classification models for regional wildlife monitoring, urban noise mapping, and localized emergency detection. These collections would improve classification accuracy and enable studies on geo-spatial acoustic patterns and culturally informed audio analytics. Collaboration among academic institutions, government bodies, and local communities is key to collecting and sharing these datasets. Ultimately, diverse and geographically tagged audio corpora would significantly increase the effectiveness and relevance of DL in site-specific audio classification tasks.

SANTRAUKA

Įvadas

Darbo aktualumas

Garso klasifikavimas atlieka svarbų vaidmenį aptinkant anomalijas, nes suteikia vertingos informacijos įvairiais taikymo atvejais [1–3]. Šios sistemos yra naudingos pramoninėse ir miesto vietovėse [4–6], taip pat sudėtingose gamtinėse vietovėse, pvz., miškuose [7]. Garso duomenys gali būti renkami neinvaziniu būdu [8, 9], tačiau kiekviena aplinka kelia unikalių iššūkių aptinkant anomalijas, ypač be vaizdo duomenų. Nors dažnai laikomas trukdžiu, foninis triukšmas yra būtinas garso informacijai užfiksuoti. Išgauti aiškų garso įrašą triukšmingoje aplinkoje yra sudėtinga, ir tai turi įtakos klasifikavimo tikslumui [10–12]. Taneja ir kt. (2013) [13] ataskaitoje nurodoma, kad tylios aplinkos įrengimas yra brangus ir nepraktiškas lauke. Žmonių, gyvūnų, mašinų ir aplinkos garsai padidina sudėtingumą, todėl sunku aptikti anomalijas [14, 15]. Šis kintamumas gali trukdyti modeliams, pakenkti duomenų kokybei ir sumažinti vertinimo tikslumą. Todėl šioms sistemoms labai svarbu naudoti signalų kondicionavimą ir savybių išskyrimą. Anomalijos šiame darbe traktuojamos kaip atskiros garso klasės arba klasių požymiai, kurių aptikimas tiesiogiai priklauso nuo klasifikavimo kokybės. Idealios klausymosi sąlygos yra retos; realaus pasaulio triukšmas daro įtaką suvokimui ir sprendimams [16–18]. Barchiesi ir kt. (2015) [19] pažymėjo, kad tokios sąlygos trikdo žmogaus suvokimą ir klasifikavimą. Naujausios inovacijos pagerino triukšmo filtravimo technologijas, padidindamos garso aiškumą ir patikimumą. Šie metodai, suderinami su dabartine technologija, pagerina pritaikomumą [20–23]. Kiekvienas garso šaltinis turi unikalų akustinį ženklą, signalizuojantį veiklą ar būseną; nukrypimai gali rodyti tokias problemas, kaip mechaniniai gedimai. Tokie patobulinimai yra svarbūs norint aptikti anomalijas triukšmingomis sąlygomis, padėdami realaus laiko stebėjimui ir automatizuotam sprendimų priėmimui. Nuolatiniai kompiuteriniai ir mokslinių tyrimų pasiekimai skatina pažangą šioje srityje.

Naujausi tyrimai atskleidžia, kad kompiuterinės technologijos, ypač AI, ML ir DL, gerokai pagerino tyrimų tikslumą be papildomų patobulinimų [24]. AI, ypač ML ir DL, gerina klasifikaciją sudėtingose aplinkose. DL modeliai efektyviai identifikuoja duomenų rinkinių modelius ir pateikia tikslias prognozes, kai tradiciniai metodai nesiekia. Sėkmė priklauso nuo kokybiškų mokymo duomenų [25]. Veiksmingam AI taikymui būtini išsamūs duomenų rinkiniai. Šios technologijos gerina garso klasifikaciją ir anomalijų nustatymą. Zada ir kt. (2022) [79] siūlo aplinkos triukšmą duomenų rinkiniuose padidinti, o Jaiswal ir Provost (2023) [80] patvirtina, kad jis padidina veiksmingumą. Ši strategija ypač naudinga sudėtingoms įgarsinimo užduotims su nesubalansuotais duomenų rinkiniais. Garso klasifikavimas apima šaltinių ir dažnių identifikavimą [26]. Pramonėje kiekvienos mašinos unikalūs garso signalai rodo problemas [27]. Gamtos garsai reikalauja kitokios strategijos: miškų garsai pateikia ekologinę akustiką [28]. Miškų analizė reikalauja pažangių DL technikų didesniam skaičiui technikų klasifikuoti [29–33]. Vis sudėtingėjančios

užduotys reikalauja pažangių modelių ir DL technikų, ypač aplinkos, aplinkos ar gyvūnų garsų klasifikavimui [30, 34–36].

Per pastaruosius dešimt metų mokslininkų komanda susitelkė į reiklią užduotį – rinkti ir vertinti garso duomenis, kad spręstų balso atpažinimo iššūkius esant foniniam triukšmui. Tikslas – sukurti duomenų rinkinius, kurie padėtų ateities tyrimams ir atspindėtų realaus pasaulio sąlygas. 2015 m. Piczak [5] pristatė naują ESC-50 duomenų rinkinį su 50 garsų iš gyvenamosios ir miesto aplinkos, įskaitant kosulį ir žingsnius, paimtus iš FreeSound.org [37]. 2019 m. Purohit ir kt. iš „Hitachi Ltd.“ ėmėsi pramoninių mašinų garsų rinkimo, sukurto MIMII rinkinio [4] su įrašais esant -6, 0 ir +6 dBs. 2023 m. Bandara ir kt. [7] išleido FSC22 rinkinį su gamtos garsais, tokiais kaip vėjas ir liūtų riaumojimas, miško ekosistemai dokumentuoti. Kiti reikšmingi rinkiniai apima UrbanSound8k [6], AudioSet [17], DCASE [38], ToyADMOS [39] ir FSD50K [40], kurie skirti įvairiems aplinkos scenarijams.

Šioje disertacijoje tiriamas atsparus, garso signalais pagrįstas anomalijų aptikimas, vertinant DNN modelių, daugiausia paremtų CNN ir transformerių architektūromis, veikimą esant įvairioms akustinėms sąlygoms. Pramoniniai, miesto ir miško garsovaizdžiai laikomi vienas kitą papildančiais domenais, aprėpiančiais struktūruotą mašinų keliamą triukšmą, įvairialypę žmogaus veiklą ir itin dinamiškus gamtinius garsus. Pasitelkiant viešai prieinamus duomenų rinkinius, tokius kaip MIMII [4], ESC-50 [5] ir FSC22 [7], nagrinėjama modelių generalizacija esant skirtingiems triukmo lygiams, klasių disbalansui ir skirtingoms spektrinėms charakteristikoms.

Perceptyviai motyvuotos požymių reprezentacijos, tokios kaip mel-spectrogram [41, 42], MFCC [42] ir chroma-STFT [43–45], naudojamos kaip įvestis hibridiniams DL modeliams. Šie modeliai sujungia iš anksto išmokytus konvoliucinius arba transformerių pagrindu sukurtus karkasus su BiLSTM ar BiGRU sluoksniais, taip vienu metu fiksuodami erdvinę ir laikinę struktūrą. Vieninga k-folds CV vertinimo procedūra užtikrina statistiškai patikimą ir tarpusavyje suderintą rezultatų analizę visiems duomenų rinkiniams. Šiame darbe anomalijos suprantamos kaip garso įvykiai ar veikimo būsenos, kurių spektrinės-laikinės savybės reikšmingai skiriasi nuo tipinių ar tikėtinų šablonų, ir jos traktuojamos kaip atskiros klasės prižiūrimos daugiaklasės klasifikacijos sistemoje.

Problemų formulavimas

1. Nepakankami ir stipriai nesubalansuoti garso duomenų rinkiniai, ypač pramoninėse anomalijų aptikimo užduotyse, mažinantys modelių patikimumą ir stabilumą.
2. Neaiški skirtingų garso požymių atvaizdavimų sąveika su šiuolaikinėmis DL architektūromis triukšmingose ir heterogeniškose akustinėse aplinkose.
3. Ribotas SNR variacijos ir duomenų papildymo metodų poveikio supratimas klasifikavimo patikimumui anomalijų aptikimo kontekste.
4. Nepakankamai ištirtas laikinio modeliavimo (pvz., GRU, LSTM) poveikis klasifikavimo tikslumui, priklausomai nuo pasirinktų požymių ir architektūrų.
5. Trūkstamas sistemingas požymių ir modelių suderinamumo vertinimas, siekiant

optimizuoti tikslumą tame pačiame duomenų rinkinyje.
6. Neužtikrintas balansas tarp klasifikavimo tikslumo, mokymo stabilumo ir skaičiavimo sąnaudų praktinėse ir ribotų resursų taikymo aplinkose.

Tyrimo tikslas

Šio tyrimo tikslas – pagerinti garso pagrindu veikiančių anomalijų aptikimo sistemų tikslumą, patikimumą ir atsparumą triukšmui, analizuojant skirtingų garso požymių atvaizdavimų ir DNN architektūrų sąveiką įvairiose akustinėse aplinkose.

Tyrimo uždaviniai

1. Išnagrinėti ankstesnius tyrimus garso pagrindu veikiančios anomalijų aptikimo ir aplinkos garsų klasifikavimo srityje, daugiausia dėmesio skiriant požymių atvaizdavimams, DL modeliams ir vertinimo metodams triukšmingose aplinkose.
2. Įvertinti duomenų disbalanso ir skirtingų triukšmo lygių poveikį anomalijų aptikimui pramoniniuose garso duomenų rinkiniuose bei išanalizuoti papildymo metodų efektyvumą.
3. Ištirti skirtingų garso požymių atvaizdavimų gebėjimą atskirti normalius ir anomalius akustinius modelius įvairiose triukšmingose aplinkose.
4. Sistemingai įvertinti skirtingų DL architektūrų (konvoliucinių, transformerių ir hibridinių laikinio modeliavimo metodų) tinkamumą anomalijų aptikimo užduotims.
5. Išanalizuoti laikinio modeliavimo strategijų poveikį anomalijų aptikimo tikslumui, atsižvelgiant į jų sąveiką su požymių atvaizdavimais ir bazinėmis architektūromis.
6. Įvertinti kompromisą tarp anomalijų aptikimo tikslumo, modelių patikimumo ir skaičiavimo sąnaudų, siekiant užtikrinti metodų tinkamumą taikant praktikoje.

Mokslinės naujovės

Šiame tyrime pristatomos šios mokslinės naujovės, kurios išplečia garso pagrindu veikiančią anomalijų aptikimo ir aplinkos garsų klasifikavimo metodologiją:
1. Sistemiškai išnaudotas visas MIMII duomenų rinkinys, apimantis visus įrenginius ir skirtingus SNR lygius, leidžiantis įvertinti modelių veikimą realiomis pramoninėmis sąlygomis ir pasiekti aukštesnį klasifikavimo tikslumą nei ankstesniuose tyrimuose [23, 46].
2. Nustatyta, kad mel-spectrogram požymių atvaizdavimas nuosekliai pranoksta tradicinius MFCC ir chroma-STFT metodus, tačiau kartu reikalauja didesnių skaičiavimo resursų, taip atskleidžiant esminį kompromisą tarp tikslumo ir skaičiavimo efektyvumo.
3. Parodyta, kad mel-spectrogram ir SWinT-BiLSTM architektūros derinys yra ypač veiksmingas riboto dydžio ir heterogeniškuose duomenų rinkiniuose, tokiuose kaip ESC-50 ir FSC22, bei išlieka konkurencingas ir sudėtingesniuose, nesubalansuotuose pramoniniuose duomenyse.
4. Įrodyta, kad pasikartojančių neuroninių tinklų sluoksnių, ypač BiLSTM, integravimas į iš anksto išmokytas architektūras padidina laikinio modeliavimo

gebėjimus ir atitolina persimokymo pradžią, nors kartu padidina modelio sudėtingumą ir treniravimo trukmę.

Praktinė reikšmė

1. Giliųjų neuroninių tinklų taikymas garso pagrindu veikiančiai anomalijų analizei sudaro prielaidas kurti neinvazines ir neardomąsias aplinkos stebėsenos sistemas, kurios gali būti diegiamos įvairiose aplinkose, nepažeidžiant vykdomų procesų.
2. Akustinių metodų ir giliojo mokymosi integracija leidžia efektyviai aptikti ir klasifikuoti reiškinius tiek pramoniniuose, tiek natūraliuose kontekstuose. Nors aplinkos triukšmas dažnai laikomas trukdžiu, tinkamai taikomi signalų apdorojimo metodai leidžia jį išnaudoti kaip papildomą informacijos šaltinį.
3. Tyrimo rezultatai rodo, kad garso požymių ir modelių parinkimas turi būti derinamas su turimais skaičiavimo resursais, siekiant praktinio pritaikomumo ir patikimo sistemų veikimo.

Disertacijos teiginiai

1. MIMII duomenų rinkiniui būdingas ryškus klasių disbalansas buvo sėkmingai sumažintas taikant SNR-pagrįstas duomenų papildymo strategijas, kurios reikšmingai pagerino modelių tikslumą, patikimumą ir apibendrinimo gebą triukšmingomis sąlygomis.
2. Tyrimas parodė, kad garso požymių atvaizdavimo pasirinkimas yra esminis veiksnys giliųjų neuroninių tinklų pagrindu veikiančiai anomalijų aptikčiai. Mel-spectrogram užtikrina didžiausią atsparumą triukšmui ir stabilumą, MFCC pasižymi geru efektyvumo ir tikslumo balansu, o chroma-STFT yra netinkamas netoninių, triukšmu dominuojamų signalų analizei.
3. Aiškiai išreikštas laikinis modeliavimas, integruojant BiGRU ir BiLSTM sluoksnius, yra būtinas siekiant patikimos anomalijų aptikimo veiklos nestacionariose ir laike kintančiose akustinėse aplinkose.
4. Siūloma klasifikavimo sistema pasižymi universalumu skirtinguose pramoniniuose, urbanistiniuose ir natūraliuose duomenų rinkiniuose, o tai patvirtina aukšti F1 rodikliai, jautrumas ir aiškus klasių atskyrimas t-SNE vizualizacijose.
5. Ankstyvo stabdymo taikymas leido efektyviai išvengti persimokymo, o rekursinių sluoksnių integravimas pagerino mokymo stabilumą ir rezultatų nuoseklumą skirtinguose duomenų rinkiniuose.

Moksliniai aprobavimai

Šio disertacinio tyrimo metodologiją ir rezultatus pagrindžia penkios mokslinės publikacijos: du straipsniai žurnaluose, indeksuojamuose *Web of Science* ir *Scopus*, bei trys publikacijos tarptautinių konferencijų leidiniuose. Išsamus publikacijų sąrašas ir jų ryšys su disertacijos turiniu pateikiami priede.

Disertacijos struktūra

Disertaciją sudaro penki skyriai. „Įvadas“ skyriuje pristatoma tyrimo problema ir jos motyvacija, apibrėžiama šiame darbe vartojama akustinių anomalijų sąvoka ir pagrindžiamas pramoninių, urbanistinių bei miško garso aplinkų pasirinkimas kaip

papildomų vertinimo sričių. „Literatūros apžvalgos apie giluminio neuroninio tinklo metodą akustinių anomalijų įvykiams aptikti“ skyriuje pateikiama išsami literatūros apžvalga, apimanti pagrindinius ir naujausius dirbtinio intelekto, mašininio ir giliojo mokymosi metodų taikymo garso atpažinimo bei klasifikavimo srityse rezultatus. Šie metodai sudaro pagrindą įvykių aptikimui, anomalijų nustatymui ir galimų grėsmių identifikavimui įvairiose aplinkose. „Duomenų rinkiniai ir metodologija“ skyriuje aprašomi tyrime naudoti duomenų rinkiniai ir eksperimentinė metodika: pristatomi pasirinkti garso rinkiniai (MIMII, ESC-50, FSC22), jų akustinės savybės ir taikomos laiko ir dažnio požymių reprezentacijos (mel-spectrogram, MFCC, chroma-STFT). Taip pat trumpai apibūdinami siūlomi hibridiniai modeliai, paremti konvoliucinėmis architektūromis ir dėmesio mechanizmus integruojančiais giliojo mokymosi modeliais, bei nurodomi validavimo protokolas, eksperimentų eiga ir vertinimo procedūros. „Našumo vertinimas ir analizė“ skyriuje pateikiama siūlomų ir bazinių giliojo mokymosi modelių eksperimentinė analizė, paremta mokymo ir validavimo rezultatais; modelių veikimas vertinamas pagal tikslumą, AUC, preciziškumą, atpažinimo pilnumą ir F1 matą. „Išvados ir ateities darbai“ skyriuje apibendrinami rezultatai, jie kritiškai įvertinami ir aptariamos tolesnių tyrimų kryptys.

Literatūros apžvalgos apie giluminio neuroninio tinklo metodą akustinių anomalijų įvykiams aptikti

Akustinių metodų naudojimas aplinkos stebėjimo ar pažeidimų nustatymo kontekste turi reikšmingą istorinį precedentą [105, 106]. Paprastai neįprastos sąlygos stebimose įrangose gali būti nustatytos pagal pagamintų akustinių signalų savybių kitimą, pavyzdžiui, dažnio ir amplitudės pokyčius. [107, 108]. Akustinio metodo pranašumas, palyginti su kitais metodais, yra tai, kad akustinio signalo savybės gali būti išgautos ir panaudotos norint aptikti gilesnius gedimus. Be to, Tagawa ir kt.(2021) [48] teigė, kad akustinių duomenų, kuriuos galima gauti tik mikrofonu ir kurie nėra destruktyvūs, surinkimas palengvina identifikavimo procesą, nesutrikdant veikiančios sistemos. Be to, Reubens ir kt.(2019) [109] pripažino, kad akustiniai metodai taip pat taikomi gyvų būtybių elgesio pokyčiams aptikti. Deja, renkant, suspaudžiant ir perduodant, visi surinkti signalai ir gauti vaizdai neišvengiamai užteršiami triukšmu, dėl ko atsiranda iškraipymų ir prarandama informacija.

Tyrimai rodo, kad triukšmas neigiamai veikia signalų apdorojimo veiklos efektyvumą. Todėl signalų triukšmo šalinimo svarba padidėjo šiuolaikinėse signalų apdorojimo sistemose, įskaitant vaizdų apdorojimo [110], kalbos atpažinimo [111] ir biomedicininių signalų apdorojimo, kuris yra labai svarbus medicininei diagnostikai [112], taikymus. Pasak Damaševičius ir kt. (2017) [113], triukšmas telekomunikacijose kenkia ryšio kanalams, dėl to sumažėja pralaidumas ir pablogėja signalo kokybė, pasireiškianti svyravimais ir informacijos praradimu. Be to, Picaut ir kt. (2020) [114] teigė, kad miesto triukšmas neigiamai veikia miesto gyventojų sveikatą ir didina triukšmo taršą. O Kantova ir kt. (2021) [115] pastebėjo, kad triukšmas kelia iššūkių daugelyje pramonės sričių ir statybos inžinerijos srityje.

Pramoninis triukšmas apibūdinamas kaip darbo vietose ir įmonėse esantis garsas, atsirandantis dėl gamybos veiklos ir mašinų, įrankių ar įrangos naudojimo [116]. Tokiam triukšmui esant, pasireiškia keletas padarinių, įskaitant pramoninės įrangos tarnavimo laiko sutrumpėjimą ir padidėjusią pramoninių avarijų riziką. Panašiai struktūriniai virpesiai panašūs į triukšmą savo potencialu pakenkti struktūrų vientisumui ir funkcionalumui. Tokie virpesiai gali sukelti įvairių neigiamų pasekmių, pvz., struktūrų nuovargio gedimus [117], nepatogumus vartotojams ar praeiviams [118], trukdžius jautriems prietaisams ir kt. [119]. Inžinerijoje pagrindinis žingsnis, būtinas būklės stebėjimo ir gedimų diagnostikos įgyvendinimui, yra išsamus vibracijos duomenų įvertinimas. Šios analizės tikslas – nustatyti svarbiausias problemines savybes, taip pagerinant diagnostikos ir vertinimų tikslumą. Todėl tiksliai analizuoti triukšmą įrašytuose vibracijos signaluose yra labai svarbu, norint tiksliai įvertinti įrenginio gedimus.

Kitaip nei pramoniniai, gyvenamieji ir miškų regionai turi unikalių akustinių savybių. Gyvenamuosiuose rajonuose daugiausia dominuoja su žmogaus veikla ir galbūt su naminiais gyvūnais susiję triukšmai [120–122]. Ši garso aplinka apima įvairius triukšmus, pvz., kasdienius pokalbius, žaidžiančius vaikus, vejapjovių gaudesį, šunų lojimą, eismo gaudesį ir netgi su dantų valymu susijusius garsus. Kita vertus, laukinės gamtos aplinkoje vyrauja natūralūs garsai, nesusiję su žmogaus veikla [123]. Šie natūralūs akustiniai elementai gali apimti vilkų kaukimą, tigrų urzgimą, liūtų riaumojimą, paukščių čiulbėjimą ir vėjo gūsius. Todėl parametrai, skirti anomalijoms šiose aplinkose identifikuoti, labai skiriasi nuo tų, kurie taikomi pramoninėse aplinkose.

Istoriškai akustiniai signalai vaidino svarbų vaidmenį aptinkant anomalijas įvairiose srityse, pvz., mechaninės diagnostikos, struktūrų būklės stebėjimo ir aplinkos stebėjimo [105, 106]. Iš pradžių akustinių anomalijų aptikimas priklausė nuo žmogaus klausos gebėjimų, operatoriai naudodavo savo klausą, kad pastebėtų garso nukrypimus, pvz., dažnio pokyčius, neįprastus rezonansus ar nereguliarias vibracijas, sukeltas mašinų ar aplinkos įrenginių. Šis metodas buvo pagrįstas plačiai pripažinta nuomone, kad sistemos akustinių savybių pokyčiai paprastai signalizuoja apie galimas problemas ar gedimus [107, 108]. Nors šis metodas buvo praktiškas ir reikalavo minimalios įrangos, jis buvo subjektyvus ir kintamas. Rezultatai labai priklausė nuo asmens kompetencijos, klausos gebėjimų ir konteksto supratimo, o tai lėmė nenuoseklumą ir ribotą atkuriamumą. Be to, žmogaus ausies ribotumas aptikti nedidelius akustinius pokyčius dažnai uždelsia ankstyvosios stadijos gedimų identifikavimą, didindamas veiklos sutrikimų ar aplinkos žalos riziką. Sistemoms tampant vis sudėtingesnėms ir reikalaujančioms nuolatinio stebėjimo realiuoju laiku, rankinio klausos vertinimo trūkumai tapo dar akivaizdesni. Ši situacija paskatino pereiti prie labiau kiekybinių ir automatizuotų metodų, naudojant signalų apdorojimą ir pažangius algoritmus akustiniams duomenims vertinti ir interpretuoti [48], taip padedant pagrindą šiuolaikinių akustinių anomalijų aptikimo sistemų kūrimui.

Sparčiai tobulėjant skaitmeniniam signalų apdorojimui, su rankiniu garsiniu

aptikimu susiję apribojimai paskatino sistemingų akustinių duomenų analizės metodų kūrimą. Tyrėjai pradėjo išgauti iš garso signalų išmatuojamas savybes, įskaitant amplitudės pokyčius, dominuojančius dažnius, spektrinę entropiją ir energijos pasiskirstymą. Šios savybės yra būtinos sistemos veikimui ir būklei įvertinti. Naudojant šiuos signalo lygio deskriptorius, neapdorotos garso bangos formos gali būti konvertuojamos į struktūrizuotus duomenis, kurie yra labiau suderinami su algoritmine analize. Šis transformavimas lėmė sistemų, galinčių identifikuoti anomalijas, nustatant nukrypimus nuo numatytų ar išmoktų signalo modelių, sukūrimą, skatinant objektyvesnį ir atkartojamą anomalijų aptikimo metodą [48]. Skirtingai nuo rankinių metodų, šios sistemos pasižymi didesniu jautrumu ir laiko skiriamąja geba, todėl galima anksti aptikti gedimus ir sumažinti priklausomybę nuo žmogaus interpretacijos. Šis pokytis žymėjo keičiamų akustinių stebėjimo sistemų, kurios palaiko automatizuotą analizę įvairiose srityse ir veiklos lygmenyse, nuo atskirų mechaninių dalių iki didelio masto aplinkos stebėjimo tinklų, eros pradžią.

ML metodų integravimas dar labiau pagerino akustinių anomalijų aptikimo galimybes, nes sistemos galėjo išmokti atpažinti išskirtinius požymius iš pažymėtų duomenų. Gerai žinomi ML klasifikatoriai, pvz., SVM, sprendimų medžiai, kNN, ir pan. buvo vieni iš pirmųjų šiam tikslui pritaikytų įrankių. Šie modeliai leido sukurti statistinius ir taisyklėmis pagrįstus modelius, galinčius atskirti normalius ir anomalinius akustinius įvykius [27, 124]. Jų veiksmingumas buvo ypač akivaizdus struktūrizuotose aplinkose, kur anomalijų tipai ir gedimų režimai buvo gerai apibrėžti ir nuosekliai pateikti mokymo duomenyse. Susiejant išgauto signalo savybes su konkrečiomis veikimo būsenomis, ML algoritmai palengvino gedimų atpažinimo automatizavimą ir sumažino priklausomybę nuo rankinio tikrinimo. Tačiau šie modeliai dažnai reikalavo išsamių savybių inžinerijos ir sunkiai susidorojo su apibendrinimu dinamiškose ar triukšmingose akustinėse aplinkose, taip sudarant sąlygas patikimesniems sprendimams, kuriuos įgalina DL.

Nuolatinis DL augimas reiškia revoliucinį pokytį akustinių anomalijų identifikavimo srityje, užtikrinantį neprilygstamą pažangą tiek tikslumo, tiek lankstumo atžvilgiu. Dabartinėje srityje DNNs, ypač struktūros, pagrįstos CNNs ir RNNs, yra nuosekliai naudojamos siekiant nuodugniai ištirti tiek neredaguotus, tiek iš anksto apdorotus garso signalus [48]. Šie pažangūs modeliai yra nepaprastai pritaikyti savarankiškai mokytis sudėtingų, hierarchinių ir abstrakčių duomenų atvaizdų, veiksmingai fiksuojant esminius laiko ir spektrinius signalus, būtinus anomalijoms aptikti. Priešingai nei tradiciniai ML metodai, DL metodai smarkiai sumažina priklausomybę nuo rankiniu būdu suprojektuotų funkcijų, palengvindami geresnį apibendrinimą net sudėtingomis ir triukšmingomis sąlygomis. Šių modelių atsparumas ir mastelio keitimas juos padarė pagrindinėmis sudedamosiomis dalimis šiuolaikinių garso aptikimo sistemų, ypač kontekstuose, kuriuose reikalingas didesnis jautrumas ir realaus laiko apdorojimo galimybės [106, 107].

Pramonės sektoriai patyrė išskirtinius privalumus dėl šių revoliucinių inovacijų. Shaikh ir kt. (2021) [27] įrodė išskirtinį giluminio mokymosi naudojimo akustinėje

diagnostikoje efektyvumą, pasiekdamas neprilygstamą tikslumą identifikuojant mechaninius gedimus, palyginti su ribotu tradicinių metodų tikslumu. Qurthobi ir kt. (2022) [124] pabrėžė transformacinį realaus laiko klasifikavimo sistemų poveikį, naudojant NN prognostinei priežiūrai ir aktyviam ankstyviam gedimų aptikimui besisukančiose mašinose su puikiu laiko ir tikslumo suderinimu. Bhuiyan ir Uddin (2023) [125] pasiūlė novatorišką multimodalinę strategiją, kuri derina akustinius ir vibracijos duomenis, smarkiai padidindama gedimų diagnostikos patikimumą sudėtingose ir reikliose mechaninėse aplinkose. Be to, siekiant ekologinės sinchronizacijos, bioakustikos taikymas tapo populiarus, nes Sharma ir kt. (2023) [126] veiksmingai naudojo giliuosius modelius rūšių specifiniams garsams identifikuoti, labai padėdami stebėti laukinę gamtą ir skatindamas gamtos apsaugos iniciatyvas. Kartu šie svarbūs tyrimai pabrėžia išskirtinį akustinių signalų lankstumą ir galią kaip svarbų įrankį svarbeiems įvykiams aptikti pramonės ir aplinkos srityse.

Nepaisant šių sėkmių, diegiant akustines anomalijų aptikimo sistemas realiomis sąlygomis vis dar kyla keletas iššūkių. Aplinkos triukšmas, duomenų disbalansas ir veiklos konteksto kintamumas toliau trukdo apibendrinimui ir patikimumui [127]. Be to, ribota prieiga prie didelių, anotuotų akustinių duomenų rinkinių tebėra kliūtis veiksmingam giliųjų modelių mokymui. Nepaisant to, naujausi pažangūs mažos galios kraštinių kompiuterinių įrenginių, pvz., įterptųjų skaitmeninių signalų procesorių ir mikrofonų, pasiekimai leido įgyvendinti realaus laiko išvadų darymą ribotose aplinkose. Žvelgiant į ateitį, svarbiausia tyrimų kryptis yra giliųjų mokymosi architektūrų optimizavimas anomalijoms aptikti nestacionariomis sąlygomis, modelių interpretavimo gerinimas ir standartizuotų etalonų kūrimas, siekiant paremti atkuriamą vertinimą. Šiai sričiai vystantis, akustinio jutimo ir pažangių algoritmų sinergija žada įvykių aptikimo revoliuciją srityje įvairiose srityse [105, 108].

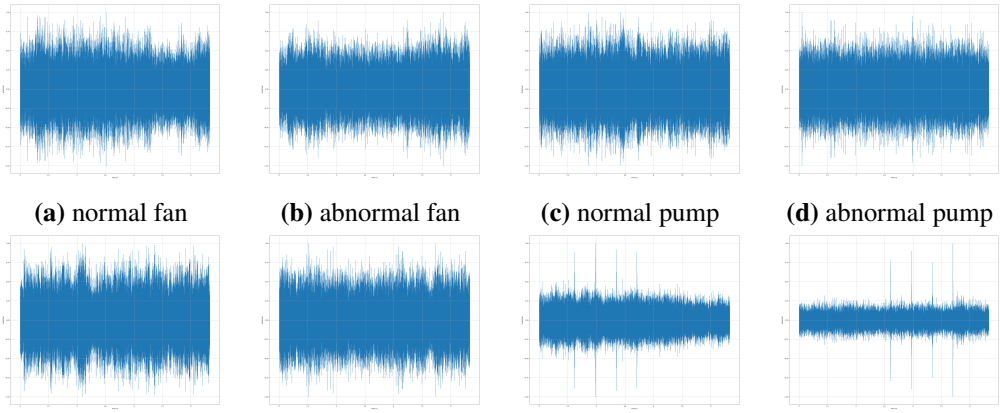
Duomenų rinkiniai ir metodologija
Duomenų minkiniai
MIMII

39 lentelė. Garso failų pasiskirstymas MIMII duomenų rinkinyje [4]

Įrenginys	ID	Sąlyga		Įrenginys	ID	Sąlyga	
		Normal	Abnormal			Normal	Abnormal
Fan	id_00	1011	407	Slider	id_00	1068	356
	id_02	1016	359		id_02	1068	267
	id_04	1033	348		id_04	534	178
	id_06	1015	361		id_06	534	89
Pump	id_00	1006	143	Valve	id_00	991	119
	id_02	1005	111		id_02	708	120
	id_04	702	100		id_04	1000	120
	id_06	1036	102		id_06	992	120

40 lentelė. Bendrosios MIMII duomenų rinkinio savybės [4]

Nuosavybė	Value
Bendras failų skaičius	∼55,000+
Failo trukmė	10 sekundės
Mėginių ėmimo dažnis	16 kHz
Kanalai	Stereo
Bitų gylis	16-bit
Failo formatas	WAV
Klasės	2 (tik operacijos)
	8 (2 (operacijos) × 4 (mašinos))
	32 (2 (operacijos) × 4 (mašinos) × 4 (identifikatoriai))
Failai pagal klasę	Skiriasi



(a) normal fan **(b)** abnormal fan **(c)** normal pump **(d)** abnormal pump

(e) normal slider **(f)** abnormal slider **(g)** normal valve **(h)** abnormal valve

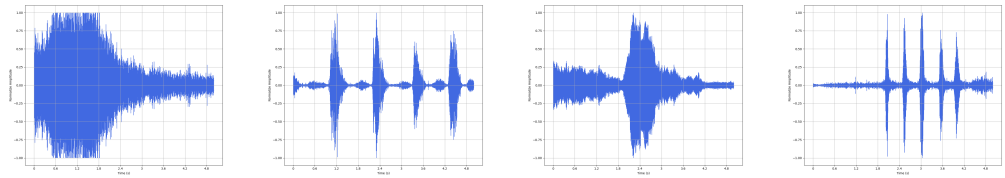
44 pav. Normalizuokite įvairių garso įrašų bangų formas MIMII duomenų rinkinyje

MIMII duomenų rinkinys apima akustinius duomenis, skirtus pramoninių mašinų būklės stebėjimo tyrimams remti. Šį rinkinį 2019 m. sudarė ir išleido Purohit ir kt. iš Hitachi Ltd R&D grupės. Jame įrašyti tipiškų gamyklos mašinų garsai realiomis sąlygomis, įskaitant ventiliatorius, siurblius, slankiklius ir vožtuvus. Jis suteikia išsamią garso informaciją apie normalią ir gedimų turinčią mašinų veiklą, padėdamas atlikti išsamią analizę. 39 ir 40 lentelės kartu su 44 pav., apibūdina įrašų pasiskirstymą ir garso charakteristikas. Šis duomenų rinkinys yra labai svarbus tyrimams, skirtiems pagerinti ML algoritmo veikimui anomalijų aptikimo srityje, kuriuos palengvina organizuoti įrašai pagal mašinų tipą, versiją ir būklę. MIMII duomenų rinkinys yra labai svarbus pramoninės diagnostikos tikslumui gerinti. Jis pripažįstamas kaip ML pažanga gedimų aptikimo srityje, nes jame yra įvairūs įrašai pagal mašinų tipus ir būsenas. Ypač, kaip matyti iš 39 lentelės, yra pavyzdžių disbalansas, kuris palankesnis normalioms būsenoms nei gedimams, o tai kelia mokymo sunkumų ir iškreipia rezultatus. Šios problemos sprendimas naudojant

duomenų strategijas yra labai svarbus norint užtikrinti patikimą modelio taikymą gal realioje aplinkoje arba realiomis sąlygomis.

ESC50

ESC-50 yra svarbus duomenų rinkinys, kuriame surinkti aplinkos garsai iš gyvenamųjų ir miesto vietovių. Pristatytas Karol J. Piczak 2015 m. [5], jis sudarytas iš Freesound.org atrinktų įrašų [37]. Kitaip nei MIMII, jame yra 2000 garsų, kurie tolygiai paskirstyti į 50 kategorijų (po 40 įrašų), užtikrinant aiškų ML ir DL modelių mokymą. Duomenų rinkinys naudoja 5 kartų *k*-folds CV sistemą, išlaikančią klasių balansą ir suteikiančią reprezentatyvius pavyzdžius [150], kas leidžia tiksliai vertinti modelius ir mažina perteklių. Piczak klasifikuoja įrašus į penkias grupes: gyvūnai (pvz., lojimas), gamta (pvz., lietus), žmogus (pvz., kosulys), vidaus (pvz., tiksėjimas) ir lauko (pvz., signalai). 45 pav. rodo pavyzdžius, 41 lentelė apibendrina savybes. Ši klasifikacija palengvina analizes kaip hierarchinė klasifikacija, didindama duomenų rinkinio suprantamumą. ESC-50 yra pagrindinė aplinkos garsų klasifikavimo priemonė, ypač DL modeliams, dėl turinio atrankos ir aktualumo. Nanni ir kt. (2021) [151] bei Zhang ir kt. (2019) [152] tyrimai patvirtina jo patikimumą.



(a) airplane **(b)** breathing **(c)** cat **(d)** dog

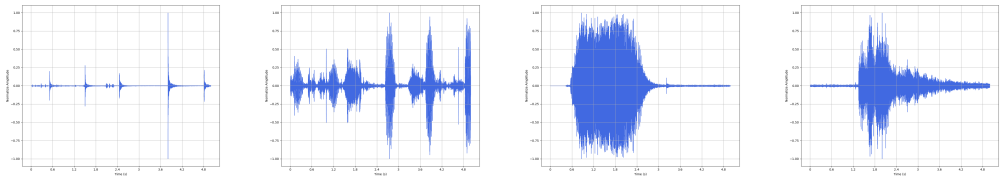
45 pav. Normalizuokite įvairių duomenų pavyzdžių bangų formas ESC-50 duomenų rinkinyje [5]

41 lentelė. ESC-50 duomenų rinkinio bendrosios savybės [5]

Nuosavybė	Value
Bendras failų skaičius	2,000
Failo trukmė	5 sekundės
Mėginių ėmimo dažnis	44,1 kHz
Bitų gylis	16-bit
Kanalai	Stereo
Failo formatas	WAV
Klasės	50
Failai pagal klasę	40

FSC22

2023 m. pristatytas FSC22 duomenų rinkinys reikšmingai prisideda prie aplinkos garsų klasifikavimo. Kitaip nei ESC-50, šis rinkinys užfiksuoja natūralius gamtos garsus su aiškiais tikslais ir unikaliu formatu, o „22“ nurodo kūrimo metus, o ne klasių skaičių. FSC22, naudodamas aukštos kokybės įrašus, atspindi įvairių miško ekosistemų garsus. Pasak Bandara ir kt. (2023) [7], jis atitinka poreikį turėti platų įrašų spektrą, kuris tiksliai reprezentuoja miško kompleksiškumą. Turinys apima 2025 anotuotus įrašus, išskirtus į 27 kategorijas: grėsmingus ir negrėsminus, palengvinančius automatizuotą stebėjimą ir ekologijos tyrimus per pažangesnius ML modelius [53]. FSC22 yra svarbus ekologinei akustikai ir grėsmių aptikimui, padedant išsaugoti gamtą realiuoju laiku aptinkant anomalijas, tokias kaip šūviai ar gyvūnų garsai [148, 153]. Natūralių ir žmogaus sukurtų garsų derinys pritaikytas įvairiems moksliniams tikslams. 46 pav. parodo bangų formų įvairovę ir pabrėžia esminį FSC22 vaidmenį būsimiems aplinkos garso tyrimams.



(a) axe **(b)** bird-chirping **(c)** speaking **(d)** thunderstorm

46 pav. Normalizuokite įvairių duomenų pavyzdžių bangų formas FSC22 duomenų rinkinyje [7]

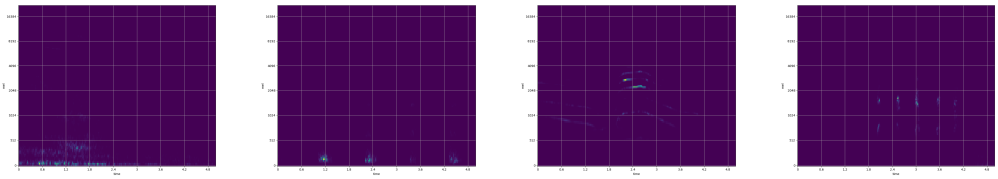
42 lentelė. FSC22 duomenų rinkinio bendrosios savybės [7]

Nuosavybė	Value
Bendras failų skaičius	2,025
Failo trukmė	5 sekundės
Mėginių ėmimo dažnis	44.1 kHz
Bitų gylis	16-bit
Kanalai	Stereo
Klasės	27
Failai pagal klasę	75

Funkcijų ekstraktoriai
Mel-spectrogram

Mel-spectrogram naudojimas yra labai svarbus apdorojant garso signalus, nes jis konvertuoja laiko srities bangų formas į suvokiamą laiko ir dažnio formatą. Skirtingai nuo įprasto STFT, kuris naudoja linijinę dažnio skalę, mel-spectrogram naudoja Mel skalę, sukurtą Stevens ir kt. 1937 m. [41], kad geriau atitiktų žmogaus klausą,

suspaudžiant aukštesnius dažnius ir išplečiant žemesnius. Tai ypač naudinga garso klasifikavimo užduotims. Plėtodami šią idėją, Davis ir Mermelstein [42] cepstralinėje analizėje įvedė Mel filtro bankus, taip sukurdami Mel pagrįstas spektrogramas. Standartinis mel-spectrogram kūrimas apima penkis etapus: STFT taikymą sudėtingai spektrogramai generuoti, jos konvertavimą į galios spektrogramą, jos atvaizdavimą Mel skalėje naudojant filtravimo bankus, pasirinktinai normalizavimą su atskaitos galia ir logaritminės kompresijos taikymą. (0.1) iki (0.6) Lygtys matematiškai apibūdina šį procesą, atsižvelgiant į spektrinį dydį ir suvokiamo garsumo pokyčius. 47 pav. iliustruoja mel-spectrogram vizualizacijas, parodydamas unikalius modelius, kurie yra būtini akustiniams įvykiams atskirti. Dėl šios savybės jie yra labai veiksmingi mokant giliųjų mokymosi modelius, skirtus aplinkos ir mašinų garsams atpažinti, ypač tais atvejais, kai labai svarbu išlaikyti laiko ir dažnio tikslumą.



(a) airplane **(b)** breathing **(c)** cat **(d)** dog

47 pav. Įvairių garso įrašų mel-spectogram vizualizacija iš ESC-50 duomenų rinkinio

$$X[t,f] = \sum_{n=0}^{N-1} x[n]\,\omega[n-tH]\,\exp\left(-j2\pi\frac{fn}{N}\right), \tag{0.1}$$

$$S_{\mathrm{P}}[t,f] = |X[t,f]|^2, \tag{0.2}$$

$$S_{\mathrm{mel}}[t,m] = \sum_{f=0}^{F-1} S_{\mathrm{P}}[t,f]\,H_m[f], \tag{0.3}$$

$$f_{\mathrm{mel}} = 2595\log_{10}\left(1+\frac{f}{700}\right), \tag{0.4}$$

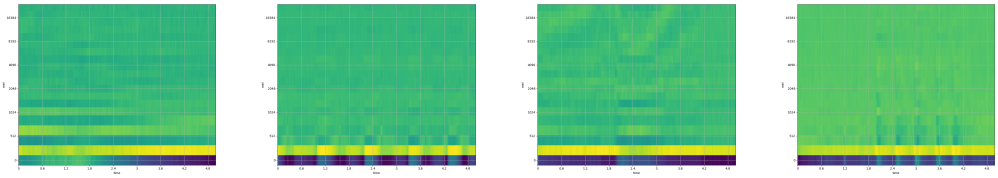
$$S_{\mathrm{mel,norm}}[t,m] = \frac{S_{\mathrm{mel}}[t,m]}{P_{\mathrm{ref}}}, \tag{0.5}$$

$$S_{\mathrm{mel,log}}[t,m] = \log\left(S_{\mathrm{mel}}[t,m]+\varepsilon\right). \tag{0.6}$$

MFCC

MFCC yra dar vienas populiarus garso savybių išskyrimui metodas, transformuojant neapdorotus bangų formos duomenis į glaustą ir suvokiamą

formatą, atitinkantį žmogaus klausą [42, 82]. Remiantis mel-spectrogram, MFCC įtraukia papildomus transformavimus, kad būtų sukurtas dekoreliuotas ir kompaktiškas savybių rinkinys. Tai apima penkis pagrindinius žingsnius: garso segmentavimą į persidengiančius kadrus, lango funkcijos taikymą, diskretišką Furjė transformavimą, kad būtų gautas galios spektras, filtravimą per Mel skalės trikampius filtrus, o tada logaritminį transformavimą, kad būtų atspindėtas žmogaus garsumo suvokimas. Paskutinis žingsnis apima naudojimą log-Mel energijoms dekoreliuoti ir suspausti savybes, išlaikant tik svarbiausius koeficientus (0.8 lygtis). Žemesnės eilės cepstraliniai koeficientai yra pabrėžiami dėl jų didesnės suvokimo svarbos, o aukštesnės eilės koeficientai, kurie paprastai fiksuoja triukšmą ar smulkias detales, dažnai yra atmesti. Ši transformacija sukuria tvirtą savybių rinkinį, naudingą klasifikavimo ir atpažinimo užduotims, ypač kai ji derinama su išankstinio apdorojimo metodais, pvz., išankstinio pabrėžimo filtravimu [160], kuris pagerina spektrinių detalių kokybę. 48 pav. pateikti MFCC atvaizdai iš ESC-50 duomenų rinkinio, kurie parodo savybių stiprumą fiksuojant platų spektrinį apvalkalą, kuris yra svarbus veiksmingai aplinkos garso analizei.



(a) airplane **(b)** breathing **(c)** cat **(d)** dog

48 pav. Įvairių garso įrašų MFCC vizualizacija iš ESC-50 duomenų rinkinio

$$X(k) = \sum_{n=0}^{N-1} x(n) \cos\left(2\pi k \frac{n}{N}\right), \tag{0.7}$$

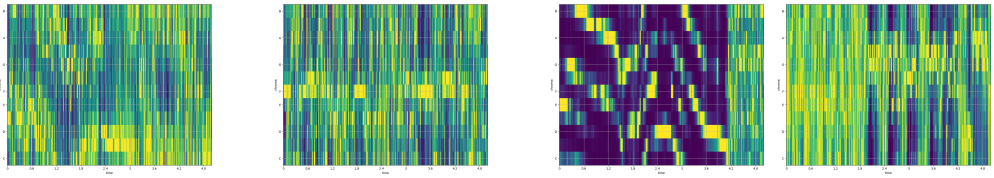
$$\mathrm{MFCC}[t,k] = \sum_{m=0}^{M-1} \log S_{\mathrm{mel}}[t,m] \cdot \cos\left(\frac{\pi k}{M}\left(m+\frac{1}{2}\right)\right). \tag{0.8}$$

Chroma-STFT

Chroma-STFT yra spektrinio atvaizdavimo technika, kurią pristatė Ellis [43] ir kuri buvo toliau pritaikyta tokioms užduotims, kaip dainų perdainavimo atpažinimas [44], tonacijos atpažinimas ir akordų įvertinimas [45, 83]. Skirtingai nuo tipinių spektrinių savybių, kurios energiją paskirsto linijiniu būdu tarp dažnių, chroma-STFT dažnius suskirsto į dvylika chromatinių tonų klasių (pvz., C, C#, D), sutelkdama dėmesį į harmoninį ir melodinį turinį, neignoruodama absoliutaus tono. Tai daro ją idealiai tinkamą muzikos srities taikymams, suderinta su vakarietiškomis toninėmis sistemomis [161, 163]. Procesas prasideda standartiniu STFT, kuris sukuria

spektrogramą, iš kurios gaunami dydžiai (0.9 lygtis). Tada dažniai susiejami su aukščio klasėmis per logaritminę skalę ir modulo-12 operaciją (0.10 lygtis), remiantis dažniu, paprastai 440 Hz (A4). Galiausiai spektrinės energijos yra surenkamos į chromos binus, sudedant kiekvienos aukščio klasės dydžius (0.11 lygtis).

Kaip parodyta 49 pav., šis metodas fiksuoja harmoninius elementus ir toninius pokyčius, kurie dažnai nepastebimi standartinėse spektrogramose. Nors iš pradžių jis buvo naudojamas muzikos analizei [162], naujesni tyrimai rodo, kad chroma-STFT galėtų būti naudingas platesnei garso klasifikacijai [164], ypač tais atvejais, kai svarbiausia yra harmoninė struktūra. Tačiau jo veiksmingumas netoninėse srityse, pvz., pramonės ar aplinkos garsuose, yra ribotas dėl mažesnio jautrumo neharmoniniams reiškiniams.



(a) airplane **(b)** breathing **(c)** cat **(d)** dog

49 pav. Įvairių garso įrašų Chroma-STFT vizualizacija iš ESC-50 duomenų rinkinio

$$S_{\text{mag}}[t,f] = |X[t,f]| = \sqrt{\text{Re}\,[X[t,f]]^2 + \text{Im}\,[X[t,f]]^2}, \tag{0.9}$$

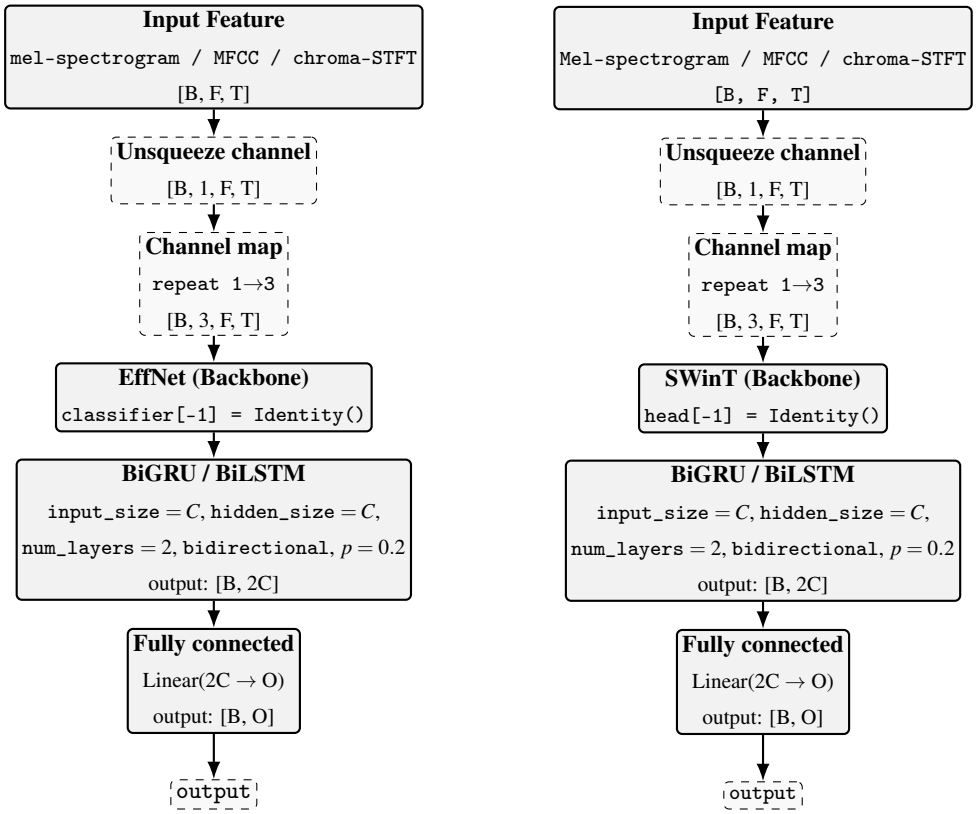
$$p(f) = \left[12 \times \log_2 \left(\frac{f}{f_{\text{ref}}} \right) \right] \mod 12, \tag{0.10}$$

$$S_{\text{chroma}}[t,p] = \sum_{f \in p} S_{\text{P}}[t,f]. \tag{0.11}$$

Siūlomi modeliai

Šiame tyrime pristatomi keli hibridiniai DL karkasai, kurie sujungia iš anksto išmokytus CNN-pagrindu sukurtus EffNet arba dėmesio mechanizmais grįstus SWinT ekstraktorius su bidirekciniais pasikartojančiais sluoksniais (BiGRU ar BiLSTM), kad būtų pagerintas laiko sekų dėsningumų suvokimas akustiniuose duomenyse, kaip pavaizduota 50 pav. Ši modulio struktūra leidžia išsaugoti erdvinį ir hierarchinį požymių mokymą pašalinus tik klasifikatorių, o rekursiniai moduliai suteikia atmintį trumpiems spektriniams pokyčiams, ypač naudinga triukšmingoje ar persidengiančių garsų aplinkoje. Architektūra yra lanksti, tinkama įvairiems iš anksto išmokytiems modeliams ir duomenų rinkiniams (MIMII, ESC-50, FSC22), o ankstesni tyrimai [165–167] patvirtino, kad RNN sluoksnių įtraukimas pagerina sekų apdorojimą ir didina klasifikacijos patikimumą. Hibridiniai variantai (i.e., EffNet-BiGRU, EffNet-BiLSTM, SWinT-BiGRU, SWinT-BiLSTM) suvienija

erdvinę-hierarchinę išraišką su laiko dimensijos tobulinimu, o panašūs tyrimai [148, 168–170] nuosekliai įrodė, kad CNN ir RNN/LSTM deriniai suteikia didelį tikslumą emocijų atpažinimo, pramoninių įrenginių būsenų ar kalbėtojų identifikacijos uždaviniuose, pasiekiant iki 94,7% ar net 90% tikslumą, taip pat pagerinant skaičiavimo našumą. Šie rezultatai rodo, kad siūlomų hibridinių modelių taikymas (50 pav.) smarkiai praplečia akustinių duomenų klasifikavimo galimybes ir sustiprina jų potencialą taikant praktiškai.

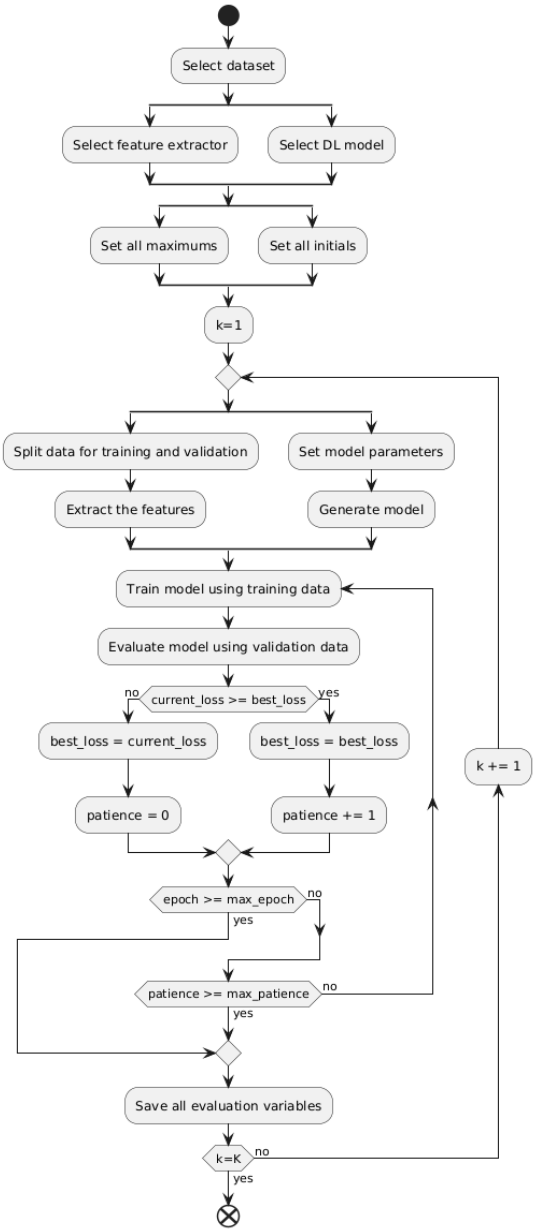


(a) EffNet-BiGRU / EffNet-BiLSTM **(b)** SWinT-BiGRU / SWinT-BiLSTM

50 pav. Siūlomi modeliai

Darbo eiga

Pateikta 51 pav. schema rodo supaprastintą šio garso klasifikavimo tyrimo procesą. Svarbiausi etapai: strateginis duomenų pasirinkimas, požymių išgavimo metodai ir DL struktūros, kurios užtikrina veiksmingą klasifikavimą ir modelių prisitaikymą įvairiose garso aplinkose. Procesas orientuotas į duomenų vientisumą ir skaičiavimo efektyvumą: prasideda nuo kruopščiai atrinktų duomenų, išgaunami esminiai požymiai, kurie apdorojami naudojant pasirinktus DL modelius. Taikomas k-folds CV, siekiant pagerinti modeliavimo apibendrinimą, ir ankstyvasis



51 pav. Darbo eigos procedūros

sustabdymas, siekiant išvengti persimokymo, taip garantuojant patikimumą.

Tinkamai parinkus duomenų rinkinius, garso požymius ir hiperparametrus, galima tiksliau įvertinti garso klasifikavimo sistemas. Šiame darbe taikomi trys pagrindiniai aplinkos garsų rinkiniai: ESC-50, MIMII ir FSC22. Esminė reikšmė tenka ir akustinių požymių atvaizdavimui, naudojant mel-spectrogram, MFCC bei

43 lentelė. Numatytųjų parametrų mel-spectrogram, MFCC ir chroma-STFT palyginimas su Librosa biblioteka

Parametras	`melspectrogram()`	`mfcc()`	`chroma_stft()`
`sr`	22050	22050	22050
`n_fft`	2048	2048	2048
`hop_length`	512	512	512
`win_length`	None	None	None
`window`	‘hann’	‘hann’	‘hann’
`center`	True	True	True
`pad_mode`	‘reflect’	‘reflect’	‘reflect’
`power`	2.0	–	–
`n_mels`	128	128	–
`fmin`	0.0	0.0	0.0
`fmax`	`sr`/2	`sr`/2	`sr`/2
`n_mfcc`	–	20	–
`n_chroma`	–	–	12
`dct_type`	–	2	–
`norm`	None	‘ortho’	None
`htk`	False	False	False
`ref`	–	1.0	–

44 lentelė. Visų parametrų ir ribinių reikšmių vertės mokymo ir validacijos metu

Parametras	**Reikšmė**	**Pastaba**
Imties dažnis	`None`	Pritaikomas prie numatytosios duomenų rinkinio vertės
Maksimalus kartų skaičius	5	Mokymo ir validacijos ciklų skaičius
Maksimalus epochų skaičius	2000	Didžiausias epochų skaičius modeliui mokyti
Maksimali kantrybė	10	Laukimo kartai prieš sustojant, jei nėra pagerėjimo
Mokymosi sparta	10^{-6}	Valdo modelio svorių atnaujinimo žingsnio dydį
Partijos dydis	32	Mėginiai apdorojami prieš svorio atnaujinimą
Nuostolių kriterijus	Kryžminė entropija	Funkcija prognozės klaidai įvertinti modelio mokymo metu
Optimizatorius	AdamW	Svorio reguliatorius nuostolių funkcijai sumažinti

chroma-STFT. `librosa` numatytieji nustatymai, apibendrinti 43 lentelėje, atskleidžia bendrus parametrus: `sr` = 22050 Hz, `n_fft` = 2048, `hop_length` = 512 ir Hann langą, leidžiančius užtikrinti nuoseklius skaičiavimus. Kiekvienas metodas turi specifines transformacijos savybes: mel-spectrogram antroji dimensija priklauso nuo `n_mels`, o

MFCC ir chroma-STFT – nuo `n_mfcc` ir `n_chroma`. Esminiai sprendimai susiję su tinkamų DL architektūrų, požymių išgavimo ir *k*-folds CV su ankstyvo sustabdymo metodu pasirinkimu. *k*-folds CV mažina šališkumą, subalansuodama duomenis mokymui ir vertinimui. Piczak (2015) [5] ir Ranmal ir kt. (2024) [143] rekomenduoja $k = 5$ ESC-50 ir FSC22 kaip efektyvų pasirinkimą. Penkių kartų kryžminė validacija suteikia gerą modelių vaizdą net ir be iš anksto nustatytų MIMII dalių. Baigiamojoje eksperimento dalyje apibrėžiami pagrindiniai hiperparametrai ir optimizavimo strategijos, lemiančios mokymo eigą ir apibendrinimą; jie stabilizuoja mokymą įvairiomis sąlygomis. Garso klasifikavimą lemia duomenų rinkinys, požymiai ir architektūra, o mokymo metu kritinį vaidmenį vaidina mokymosi iteracijos, epochos, ankstyvas sustabdymas, mokymosi sparta, partijos dydis, optimizatorius ir nuostolių funkcija, kurie veikia gradientus, derindami greitį ir stabilumą. Sistemoje taikomi adaptyvūs metodai, tokie kaip AdamW, ir kryžminės entropijos nuostolių funkcija, siekiant patikimo mokymo su skirtingais požymiais ir rinkiniais. Parametrų parinkimas grindžiamas pirminiais bandymais ir naujausiais darbais; 44 lentelėje pateikiamos jų reikšmės ir ribos, būtinos mokymo ir validacijos procesų skaidrumui ir atkuriamumui.

Našumo vertinimas ir analizė

Klasifikavimo rezultatai ir atmetimo (angl. *ablation*) tyrimai

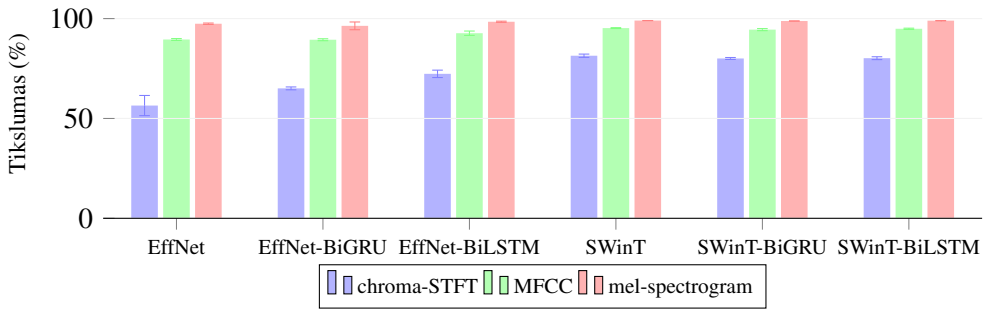
Šiame skyriuje bus pateikta įgyvendinimo rezultatų lyginamoji statistinė analizė, taikant skirtingus modelius ir požymių ištraukimo metodus. Vertinimui naudojami tokie rodikliai, kaip tikslumas, AUC, preciziškumas, atsiminimas, F1 bei t-SNE vaizdavimas. Taip pat bus nagrinėjama per didelio pritaikymo (angl. *overfitting*) problema.

Tikslumas

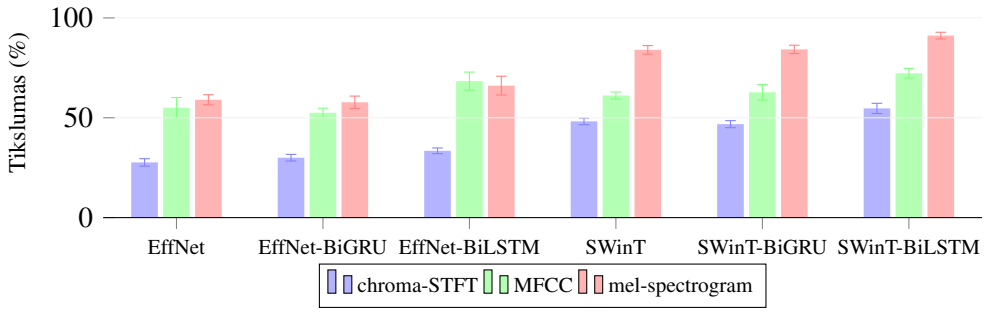
52, 53 ir 54 pav. pateikti rezultatai aiškiai parodo, kad požymių pasirinkimas ir tinklo architektūra reikšmingai lemia klasifikavimo tikslumą, taikant MIMII, ESC-50 ir FSC22 rinkinius. Visuose trijuose rinkiniuose mel-spectrogram pasirodo kaip efektyviausias požymių atvaizdavimas: jis užtikrina aukščiausią tikslumą ir mažą rezultatų kintamumą, ypač derinamas su transformerių pagrindu sukurtais modeliais. MIMII rinkinyje geriausias vidutinis tikslumas gautas naudojant bazinį SWinT modelį su mel-spectrogram – $99,02\% \pm 0,02\%$, o hibridiniai transformerių modeliai (SWinT-BiGRU, SWinT-BiLSTM) pasiekia beveik identiškus rezultatus.

ESC-50 rinkinyje taip pat dominuoja mel-spectrogram: SWinT-BiLSTM modelis pasiekia didžiausią tikslumą – $91,15\% \pm 1,67\%$, aplenkdamas tiek CNN, tiek bazinį transformerio modelį. MFCC požymiai čia duoda vidutinį našumą, o rekursinių sluoksnių integravimas (EffNet-BiLSTM, SWinT-BiLSTM) nuosekliai pagerina rezultatus, palyginti su bazinėmis versijomis.

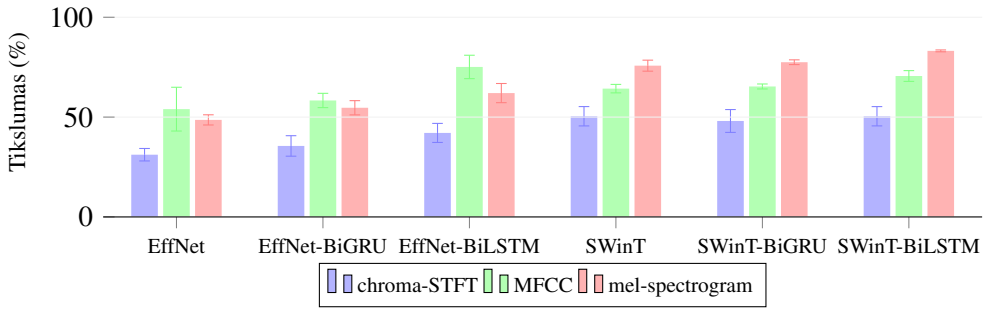
FSC22 rinkinyje bendras tikslumas mažesnis dėl didesnio akustinio sudėtingumo, tačiau išlieka tos pačios tendencijos: SWinT-BiLSTM su mel-spectrogram pasiekia didžiausią tikslumą – $83,16\% \pm 0,47\%$ ir pasižymi



52 pav. Lyginamoji statistinė klasifikavimo tikslumų analizė MIMII duomenų rinkinyje



53 pav. Lyginamoji statistinė klasifikavimo tikslumų analizė ESC-50 duomenų rinkinyje

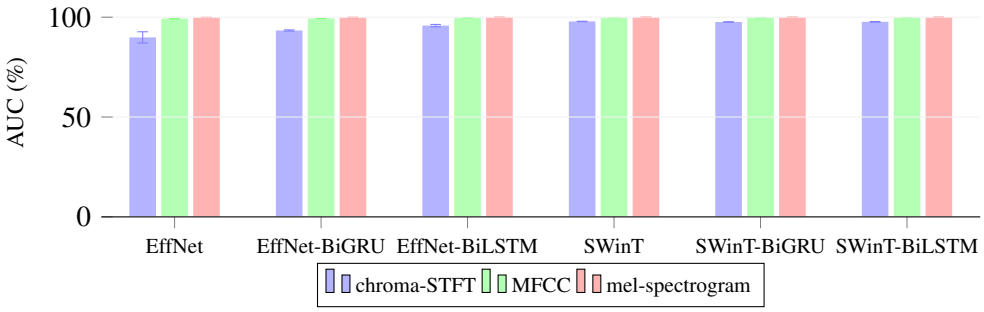


54 pav. Lyginamoji statistinė klasifikavimo tikslumų analizė FSC22 duomenų rinkinyje

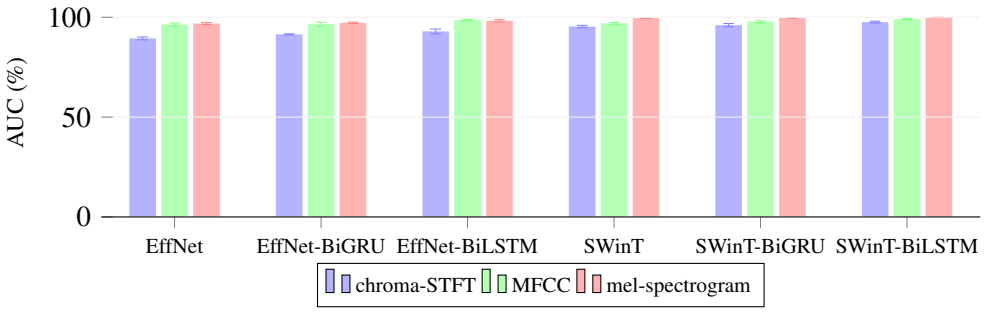
didžiausiu stabilumu. MFCC čia užtikrina konkurencingus, bet labiau svyruojančius rezultatus, o chroma-STFT visais atvejais veikia pras-čiausiai, ypač nemuzikinėse ir triukšmingose aplinkose.

Taigi grafikai patvirtina, kad pažangūs laiko ir dažnio požymiai, pirmiausia mel-spectrogram, derinami su architektūromis, turinčiomis aiškų laikinį modeliavimą ir dėmesio mechanizmus, yra kritiški siekiant aukšto ir stabilaus klasifikavimo tikslumo skirtinguose aplinkos garsų rinkiniuose.

AUC

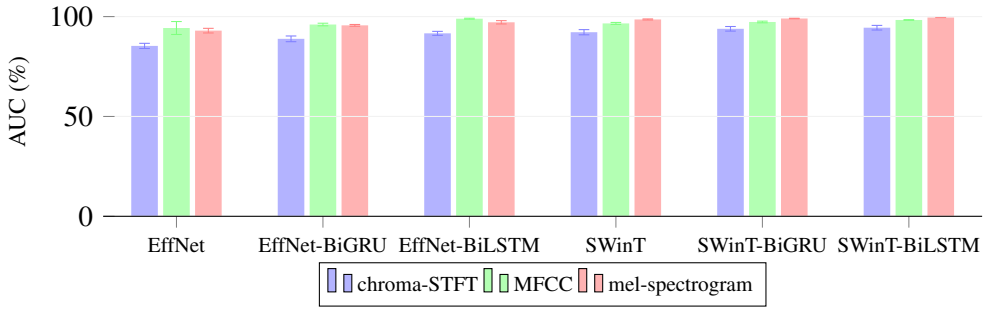


55 pav. Lyginamoji klasifikacijos AUC statistinė analizė MIMII duomenų rinkinyje



56 pav. Lyginamoji 150 AUC statistinė analizė ESC-50 duomenų rinkinyje

55, 56 ir 57 pav. pateikti AUC rezultatai parodo modelių gebėjimą atskirti klases įvairiuose akustiniuose kontekstuose. Visuose trijuose duomenų rinkiniuose aukščiausios AUC reikšmės gaunamos naudojant mel-spectrogram, ypač ją derinant su transformerių pagrindu sukurtais modeliais. MIMII rinkinyje SWinT, SWinT-BiGRU ir SWinT-BiLSTM beveik tobulai diskriminuoja klases, pasiekdami AUC iki $99,99\% \pm 0,01\%$. ESC-50 rinkinyje geriausias AUC užfiksuotas SWinT-BiLSTM modelio su mel-spectrogram – $99,93\% \pm 0,03\%$, o FSC22 rinkinyje tas pats modelis pasiekia $99,55\% \pm 0,04\%$ ir pasižymi didžiausiu rezultatų stabilumu. Priešingai, chroma-STFT visais atvejais duoda prasčiausius AUC rodiklius, o MFCC užima vidurinę poziciją, kurios našumas pagerėja pridėjus laiko modeliavimo sluoksnius. Taigi, šie rezultatai patvirtina, kad pažangūs laiko ir dažnio požymiai ir dėmesio



57 pav. Lyginamoji klasifikacijos AUC statistinė analizė FSC22 duomenų rinkinyje

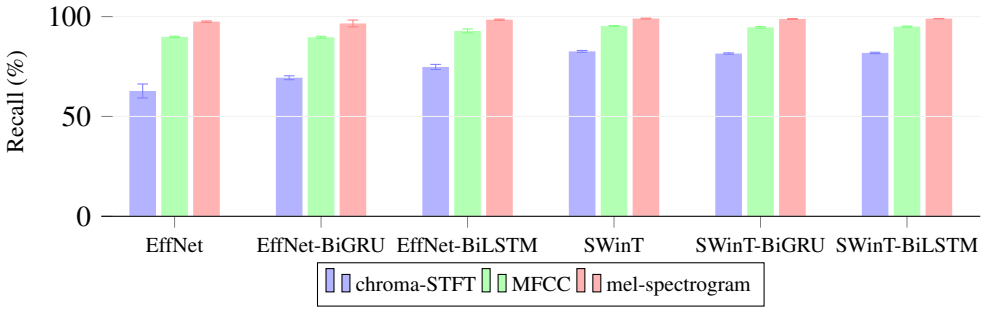
mechanizmus naudojantys modeliai yra lemiami siekiant aukšto klasifikavimo tikslumo skirtingose akustinėse aplinkose.

Preciziškumas

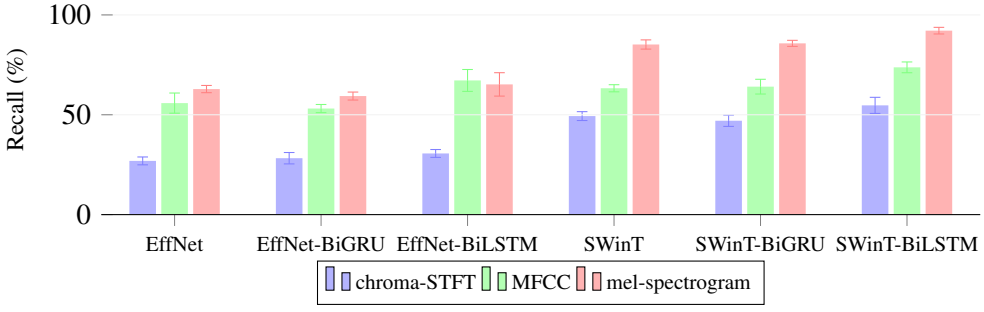
58, 59 ir 60 pav. pateikiami modelių atpažinimo preciziškumo (angl. *precision*) rezultatai MIMII, ESC-50 ir FSC22 duomenų rinkiniuose. Visuose rinkiniuose mel-spectrogram nuosekliai užtikrina didžiausią preciziškumą, ypač naudojant transformerių pagrindu sukurtas architektūras: MIMII rinkinyje geriausią preciziškumą (angl. *precision*) pasiekia SWinT ($99,03\%$), o ESC-50 ir FSC22 rinkiniuose aukščiausi rezultatai gaunami su SWinT-BiLSTM – atitinkamai $92,12\%$ ir $83,36\%$. MFCC suteikia vidutinį veikimą, kuris daugeliu atvejų pagerėja įtraukus rekursinius sluoksnius (EffNet-BiLSTM, SWinT-BiLSTM). O chroma-STFT visuose uždaviniuose išlieka silpniausiu požymių atvaizdavimu, konvoliuciniai modeliai, naudojantys šį požymį, dažnai pasižymi mažesniu preciziškumu (angl. *precision*), ypač ESC-50 ir FSC22 rinkiniuose. Apskritai rezultatai parodo, kad pažangūs laiko ir dažnio požymiai ir architektūros su dėmesio bei laiko modeliavimo mechanizmais yra kritiškai svarbūs siekiant aukšto preciziškumo (angl. *precision*) įvairiose akustinėse aplinkose.

Atpažinimo rodiklis (angl. *recall*)

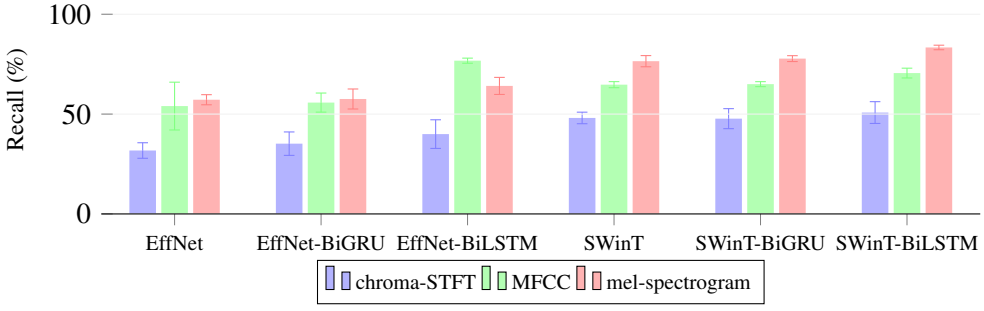
61, 62 ir 63 pav. pateikiama atpažinimo rodiklio (angl. *recall*) palyginamoji analizė, taikant skirtingas DL architektūras duomenų rinkiniuose MIMII, ESC-50 ir FSC22. Visuose rinkiniuose nuosekliai geriausi *recall* rezultatai gaunami naudojant mel-spectrogram požymius, ypač derinant juos su SWinT-pagrįstomis architektūromis ir rekursiniais sluoksniais. Konkrečiai, SWinT-BiLSTM kartu su mel-spectrogram pasiekia $98,97\% \pm 0,09\%$ MIMII, $91,15\% \pm 1,67\%$ ESC-50 ir $83,16\% \pm 0,47\%$ FSC22 rinkiniuose. MFCC požymiai išlieka konkurencingi, tačiau jų atpažinimo rodiklis mažesnis, ypač EffNet-BiLSTM ir SWinT-BiLSTM modeliuose, o chroma-STFT visuose rinkiniuose rodo žemiausius atpažinimo rodiklius, net ir



58 pav. Lyginamoji klasifikacijos preciziškumo (angl. *precision*) statistinė analizė MIMII duomenų rinkinyje

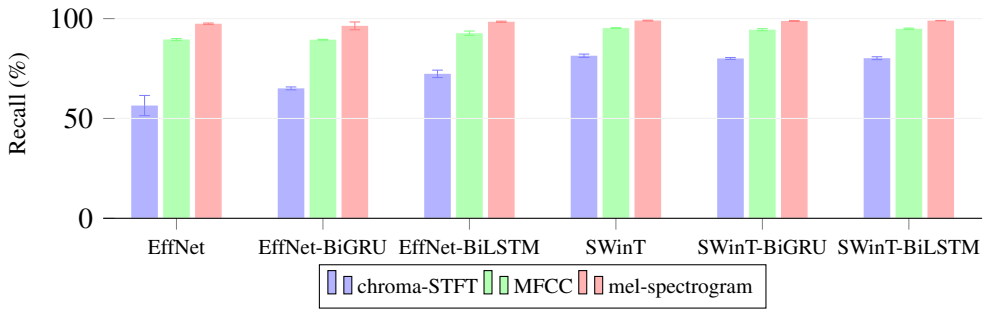


59 pav. Lyginamoji klasifikacijos preciziškumo (angl. *precision*) statistinė analizė ESC-50 duomenų rinkinyje

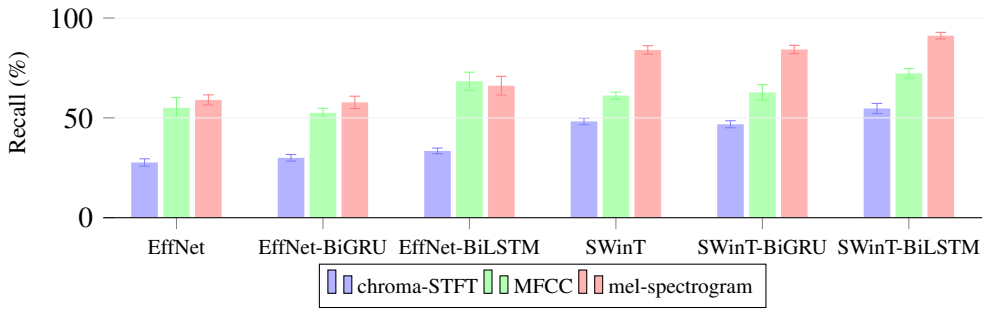


60 pav. Lyginamoji klasifikacijos preciziškumo (angl. *precision*) statistinė analizė FSC22 duomenų rinkinyje

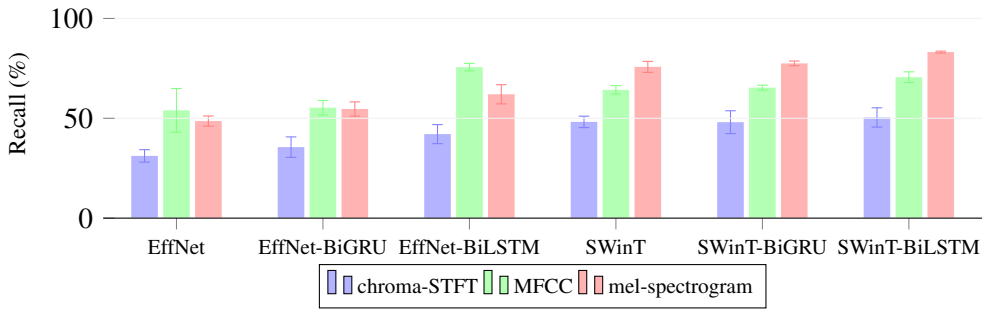
naudojant BiGRU ar BiLSTM sluoksnius. Šie rezultatai rodo, kad išsamesnė laiko–dažnio reprezentacija kartu su ilgalaikes priklausomybes modeliuojančiais hibridiniais tinklais yra kritiškai svarbi siekiant aukšto *recall* sudėtingose akustinėse



61 pav. Lyginamoji klasifikacijos atpažinimo rodiklio (angl. *recall*) statistinė analizė MIMII duomenų rinkinyje



62 pav. Lyginamoji klasifikacijos atpažinimo rodiklio (angl. *recall*) statistinė analizė ESC-50 duomenų rinkinyje



63 pav. Lyginamoji klasifikacijos atpažinimo rodiklio (angl. *recall*) statistinė analizė FSC22 duomenų rinkinyje

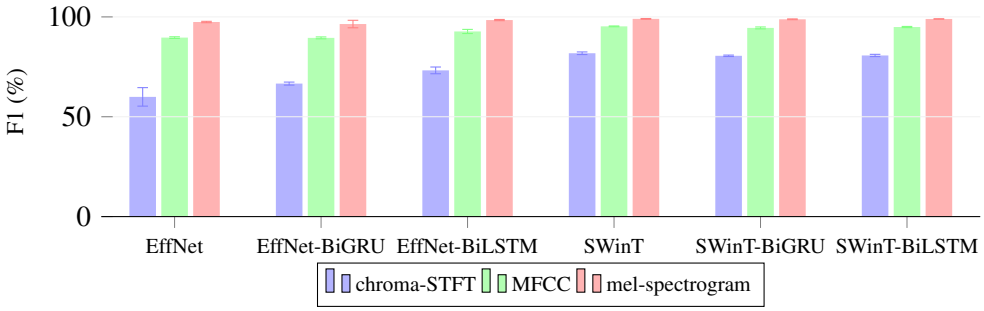
aplinkose.

F1 įvertis

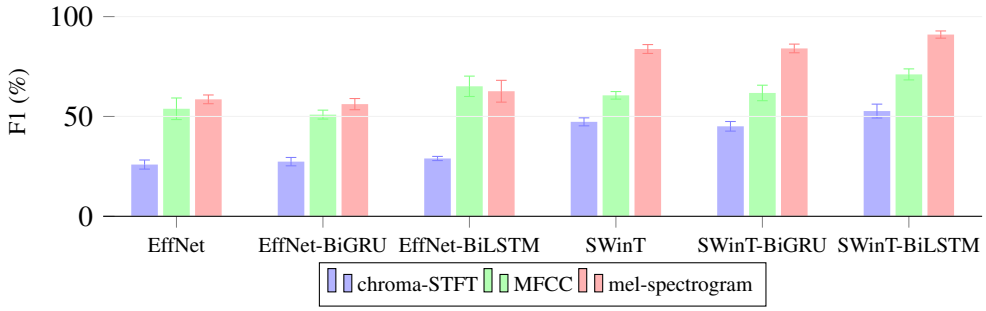
64, 65 ir 66 pav. pateikta *F1* įverčių analizė trijuose garso klasifikacijos rinkiniuose parodo nuoseklią priklausomybę tarp naudojamų požymių rinkinių ir pasirinktos modelio architektūros. Visuose trijuose rinkiniuose aukščiausi *F1* įverčiai gaunami naudojant mel-spectrogram požymius; ypač jie išryškėja derinant šią reprezentaciją su SWinT-pagrįstomis architektūromis ir rekursiniais sluoksniais, kurie leidžia efektyviau modeliuoti sekų duomenis. Geriausi rezultatai pasiekiami SWinT-BiLSTM modeliu, kurio *F1* įverčiai siekia $98,97\% \pm 0,09\%$ MIMII rinkinyje, $91,01\% \pm 1,81\%$ ESC-50 rinkinyje ir $82,47\% \pm 0,98\%$ FSC22 rinkinyje, taip pabrėžiant šios architektūros pranašumą, palyginti su kitomis nagrinėtomis alternatyvomis.

O MFCC požymiai suteikia vidutinio lygio, bet palyginti stabilius *F1* įverčius, ypač taikant EffNet-BiLSTM architektūrą, kurioje sujungiami lengvesni konvoliuciniai blokai ir rekursiniai sluoksniai. Priešingai, chroma-STFT požymių naudojimas visuose rinkiniuose duoda nuosekliai žemiausius rezultatus, net ir pritaikius sudėtingesnius rekursinius sluoksnius, tokius kaip BiGRU ar BiLSTM. Tai rodo, kad šio tipo požymiai yra mažiau tinkami nagrinėjamoms triukšmingoms ar semantiškai įvairesnėms garso atpažinimo užduotims.

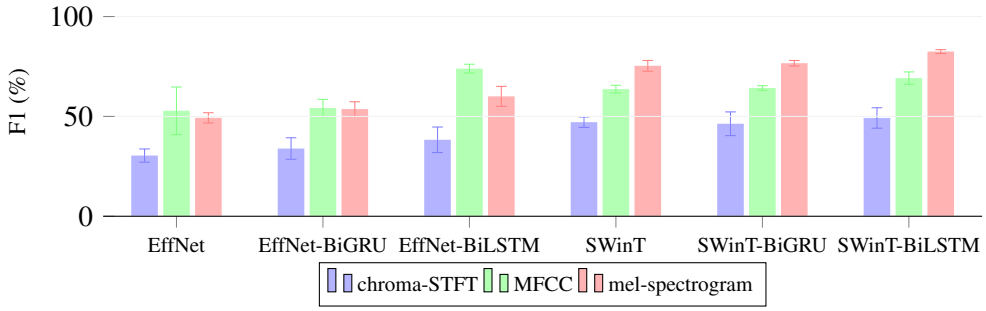
Apibendrinant, gauti rezultatai patvirtina, kad geriausias kompromisas tarp tikslumo ir bendrų atpažinimo rodiklių pasiekiamas tada, kai turtinga laiko–dažnio reprezentacija (mel-spectrogram) derinama su hibridinėmis architektūromis, gebančiomis modeliuoti ilgalaikes priklausomybes (pvz., SWinT-BiLSTM ar išplėstinės LSTM-pagrįstos struktūros). Toks derinys leidžia ne tik pagerinti klasifikavimo kokybę skirtinguose rinkiniuose, bet ir užtikrinti didesnį sprendimų stabilumą, palyginti su paprastesniais požymių rinkiniais ar architektūromis.



64 pav. Lyginamoji klasifikacijos *F1* įverčio statistinė analizė MIMII duomenų rinkinyje



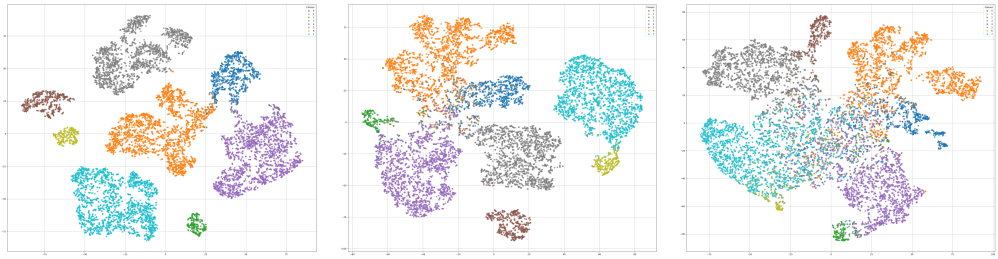
65 pav. Lyginamoji klasifikacijos *F1* įverčio statistinė analizė ESC-50 duomenų rinkinyje



66 pav. Lyginamoji klasifikacijos *F1* įverčio statistinė analizė FSC22 duomenų rinkinyje

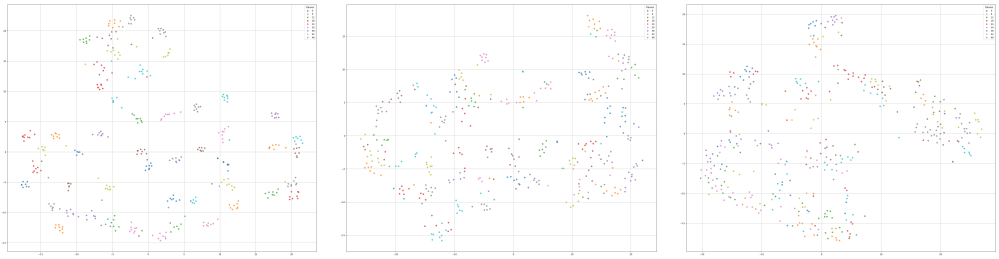
t-SNE

Vizualizacijos, tokios kaip t-SNE yra svarbios papildant įprastus našumo rodiklius vertinant garso klasifikacijos sistemas su MIMII, ESC-50 ir FSC22 duomenimis. 67, 68 ir 69 pav. matyti, kaip t-SNE perkelia kompleksinius požymius į dvimačius plotus, leidžiant lengviau stebėti klasterius ir jų atskirtis, gerinant interpretaciją [48, 182]. Vizualizacijos rodo, kad mel-spectrogram požymiai sudaro aiškius klasterius, išsaugodami svarbią spektrinę ir laikinę informaciją, būtiną sudėtingiems garsams, tokiems kaip pramoniniai gedimai ar gamtos triukšmas, atpažinti. MFCC požymiai sudaro kompaktiškus, dalinai susiliejančius klasterius, derindami efektyvumą su svarbia suvokimo informacija, bet teikdami mažiau laikinės informacijos. Chroma-STFT klasteriai yra neaiškūs ir supainioti, rodantys netinkamumą netoniniams garsams. Šios vizualizacijos paremia ankstesnes rodiklių išvadas ir parodo, kaip svarbu parinkti tinkamas garso požymių reprezentacijas pagal akustines domeno savybes. Visa tai rodo, kad t-SNE vizualizacijos padeda geriau suprasti klasifikacijos ribas ir kad teisingas požymių pasirinkimas svarbus gerinant modelių efektyvumą su įvairiais garso atpažinimo atvejais.



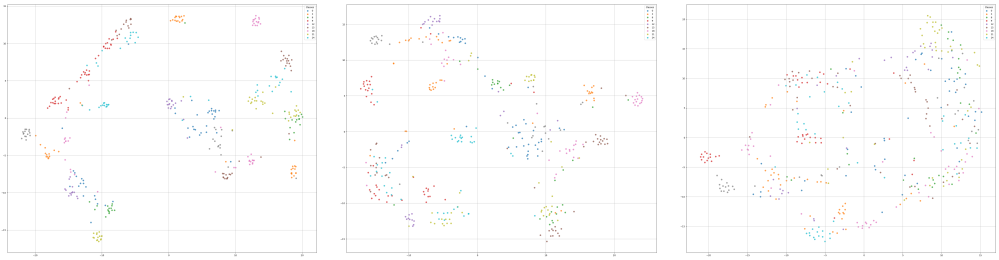
(a) mel-spectrogram **(b)** MFCC **(c)** chroma-STFT

67 pav. t-SNE vizualizacija, pasiekianti aukščiausius rezultatus klasifikuojant MIMII duomenų rinkinį naudojant įvairius požymius



(a) mel-spectrogram **(b)** MFCC **(c)** chroma-STFT

68 pav. t-SNE vizualizacija, pasiekianti aukščiausius rezultatus klasifikuojant ESC-50 duomenų rinkinį naudojant įvairius požymius

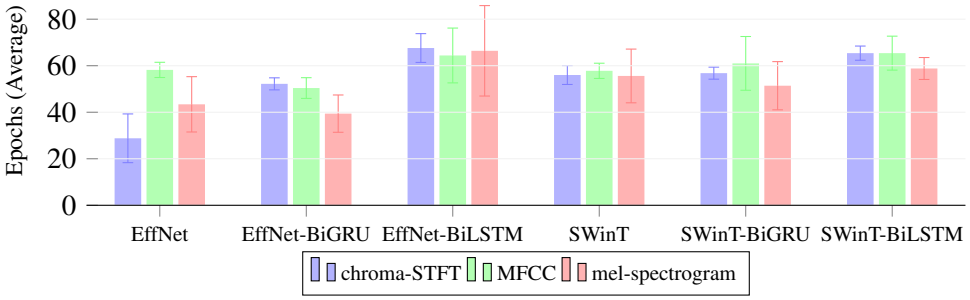


(a) mel-spectrogram **(b)** MFCC **(c)** chroma-STFT

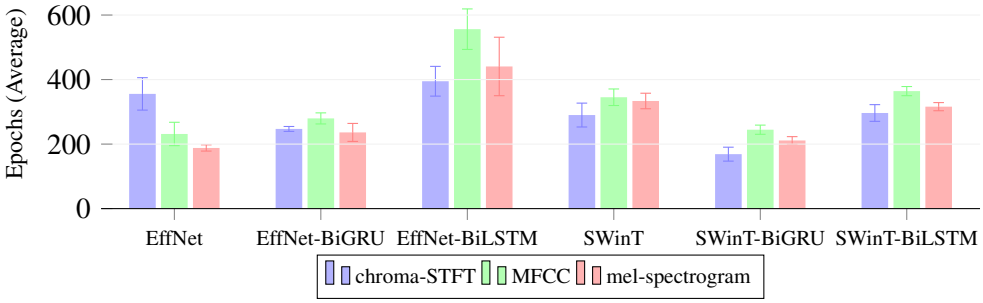
69 pav. t-SNE vizualizacija, pasiekianti aukščiausius rezultatus klasifikuojant FSC22 duomenų rinkinį naudojant įvairius požymius

Polinkis į persimokymą

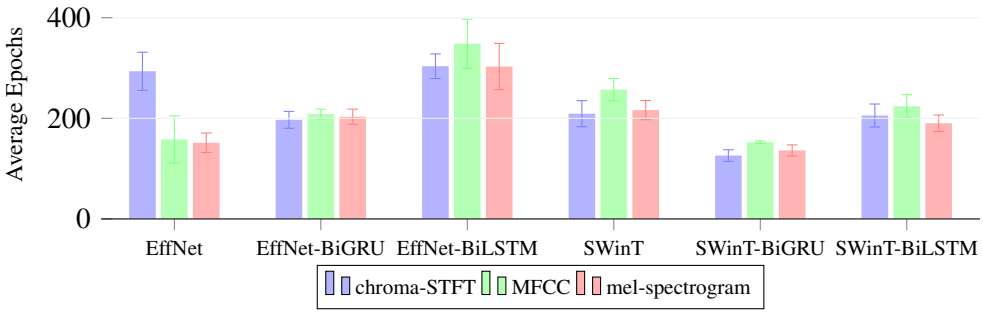
Tyrimas parodė, kad epochų skaičius iki ankstyvo sustabdymo atspindi mokymo stabilumą ir galimą persimokymo riziką skirtinguose duomenų rinkiniuose MIMII, ESC-50 ir FSC22. MIMII atveju (70 pav.) mokymas dažniausiai baigiasi gerokai



70 pav. Lyginamoji statistinė klasifikacijos epochų skaičiaus analizė su MIMII duomenų rinkiniu



71 pav. Lyginamoji statistinė klasifikacijos epochų skaičiaus analizė su ESC-50 duomenų rinkiniu



72 pav. Lyginamoji statistinė klasifikacijos epochų skaičiaus analizė su FSC22 duomenų rinkiniu

anksčiau nei 100 epochų: mažiausias vidurkis fiksuojamas EffNet su chroma-STFT ($28,80 \pm 10,47$), o ilgesnės treniruotės reikalingos gilesniems modeliams, pvz., EffNet-BiLSTM su chroma-STFT ($67,60 \pm 6,19$) ir SWinT-BiLSTM su MFCC

$(65,40 \pm 7,30)$. O ESC-50 rinkinyje (71 pav.) epochų skaičiai yra kelis kartus didesni, o didžiausias vidurkis pasiekiamas EffNet-BiLSTM su MFCC $(556,20 \pm 62,79)$, rodantis, kad sudėtingesni modeliai ir požymiai reikalauja ilgesnės konvergencijos. Greičiausiai šiame rinkinyje konverguoja SWinT-BiGRU su chroma-STFT $(168,80 \pm 21,58)$ ir EffNet su mel-spectrogram $(188,00 \pm 9,67)$. FSC22 rinkinyje (72 pav.) pastebima panaši tendencija: daugiausia epochų reikalauja EffNet-BiLSTM su MFCC $(348,20 \pm 48,82)$, o mažiausiai – SWinT-BiGRU su chroma-STFT $(126,00 \pm 11.58)$. Taigi, rekursiniais sluoksniais sustiprintos architektūros (pvz., EffNet-BiLSTM) dažniau mokomos ilgiau, ypač su MFCC, o SWinT-BiGRU nuosekliai pasiekia ankstyvą sustabdymą su mažesniu epochų skaičiumi, kas rodo efektyvesnę konvergenciją esant skirtingiems požymiams.

Diskusija

Rezultatų analizė pagal vertinimo metrikas

Rezultatai, gauti su MIMII, ESC-50 ir FSC22 duomenų rinkiniais, atskleidžia, kad aukštą klasifikavimo tikslumą užtikrina tinkamai parinktas informatyvių spektrinių požymių ir laiko priklausomybes modeliuojančių architektūrų derinys. Paprastai geriausi rodikliai pasiekiami naudojant mel-spectrograms, ypač tada, kai jos derinamos su transformerių tipo modeliais, tokiais kaip SWinT.

Priešingai, chroma-STFT pagrįsti požymiai dažniausiai generuoja mažesnį tikslumą, nes jų spektrinė raiška yra ribota, ypač triukšmingoje aplinkoje arba esant netoniniams garso signalams, kur reikalingas detalesnis dažnių struktūros atvaizdavimas. Tokiose situacijose mel-spectrograms geriau išryškina svarbias spektrines charakteristikas, kurios yra esminės klasifikavimo uždaviniams.

Nors sudėtingesnės hibridinės architektūros, derinančios kelis požymių tipus ir skirtingus sekų apdorojimo modulius, paprastai leidžia pasiekti aukštesnes kokybines metrikas, jos taip pat reikalauja gerokai daugiau skaičiavimo resursų (atminties, procesoriaus ar GPU laiko) ir ilgesnio mokymo bei inferencijos laiko. Dėl to taikant praktikoje, ypač resursų ribojamose sistemose, tampa būtina sąmoningai balansuoti tarp siekiamo tikslumo ir skaičiavimo efektyvumo, pasirenkant tokias architektūras ir požymius, kurie užtikrintų optimalų našumo ir išteklių sąnaudų santykį.

Našumo pokyčiai integruojant RNN-pagrįstus modelius

Įtraukus RNN sluoksnius, tokius kaip BiGRU ar BiLSTM, gaunami nevienodi rezultatai, kurių pobūdis priklauso nuo pasirinktos bazinės architektūros ir naudojamo požymių rinkinio. Transformerių pagrindu sukonstruotuose modeliuose papildomi rekursiniai sluoksniai dažniausiai nesuteikia apčiuopiamos naudos ir kai kuriais atvejais net sumažina našumą, nes savidėmesio mechanizmai jau pakankamai gerai įveikia ilgalaikių priklausomybių modeliavimą. Priešingai, EffNet architektūra paremtuose modeliuose RNN sluoksnių integravimas nuosekliai pagerina rezultatus, ypač tada, kai naudojami kompaktiškesni, mažiau informatyvūs požymiai, tokie kaip MFCC ar chroma-STFT. Tai leidžia daryti prielaidą, kad RNN sluoksniai yra ypač

naudingi situacijose, kai požymių reprezentacija yra ribotesnė ir mažiau išraiškinga, todėl papildomas sekų modeliavimas padeda kompensuoti šį trūkumą. Toks elgesio dėsningumas dera su Andayani ir kt. (2022) [183] pateiktais empiriniais rezultatais.

Skirtingų požymių pasirinkimo poveikis

Požymių pasirinkimas turi lemiamą įtaką klasifikavimo rezultatams visuose nagrinėtuose duomenų rinkiniuose. mel-spectrograms nuosekliai pranoksta MFCC ir chroma-STFT, ypač sudėtingose aplinkos garsų atpažinimo užduotyse, tokiose kaip ESC-50 ir FSC22. MFCC užima tarpinę poziciją – jie suteikia patrauklų kompromisą tarp klasifikavimo tikslumo ir skaičiavimo kaštų, todėl yra praktiškas pasirinkimas ribotų resursų scenarijuose. O chroma-STFT pasirodo prasčiausiai bendro pobūdžio aplinkos garsų atveju, nes šis požymių tipas iš esmės sukurtas harmoninei ir toninei struktūrai analizuoti, o ne įvairiapusėms, triukšmingoms garso scenoms [157, 185, 186]. Tokie rezultatai atitinka platesnę šiuolaikinę tendenciją giliuosiuose neuroniniuose tinkluose naudoti aukštos raiškos laiko ir dažnio reprezentacijas, kurios leidžia tinklams išgauti smulkesnius spektro ir laiko raidos niuansus bei pagerinti modelių bendrąjį apibendrinimą [86, 88, 189–191].

Epochos ir konvergencijos analizė

Epochos analizė atskleidžia kompromisą tarp greito konvergavimo ir pavojaus per anksti nutraukti mokymą. Modeliai, kuriems pakanka mažesnio epochų skaičiaus, pasiekia gerus rezultatus greičiau, tačiau tai gali reikšti ribotas galimybes išmokti sudėtingesnius duomenų dėsningumus. Priešingai, modeliai su BiLSTM sluoksniais paprastai treniruojami daugiau epochų, ypač kai naudojami MFCC ar mel-spectrogram požymiai, nes ilgesnis mokymas leidžia geriau išnaudoti laiko sekų priklausomybes ir detaliau modeliuoti signalą. Vis dėlto tai susiję su didesnėmis skaičiavimo sąnaudomis ir ilgesniu mokymo laiku. Dėl šios priežasties epochų skaičių tikslinga vertinti ne izoliuotai, o kartu su pasiekta klasifikavimo kokybe, modelio architektūros sudėtingumu ir turimais skaičiavimo ištekliais.

Palyginimas su ankstesniais tyrimais

45, 46 ir 47 lentelėse pateikiama naujausių MIMII, ESC-50 ir FSC22 klasifikavimo tyrimų, pasiekusių geriausius rezultatus, palyginamoji apžvalga. Jos suteikia kontekstą, reikalingą šiame darbe gautiems rezultatams vertinti esamų pažangiausių metodų fone.

45 lentelė parodo, kad dauguma ankstesnių darbų vertino modelių našumą gana ribotomis sąlygomis, pvz., analizuojant tik tam tikras įrenginių klases [192, 194, 196], apsiribojant dvejetaine klasifikacija [195] arba vertinant tik kelis SNR lygius [23, 46]. Nors tokiuose scenarijuose gauti rezultatai yra konkurencingi, jie sunkina tiesioginį palyginimą. Šiame darbe, priešingai, naudojant SWinT su mel-spectrograms visam MIMII rinkiniui ir vieningą k-folds CV schemą, pasiektas

45 lentelė. Naujausių tyrimų didžiausių pasiekimų palyginimas MIMII klasifikavimo uždavinyje

Autorius(-iai)	Metodas	Didžiausias tikslumas	Pastabos
Ding et al. (2023) [192]	CNN-pagrįstas modelis	94,25%	Klasifikacija atliekama atskirai kiekvienai įrenginio veikimo būsenai
Siraj et al. (2023) [193]	Few-shot mokymasis, MobileNet ir STFT	86,44%	Dėmesys skiriamas anomaliniams įrašams, naudojami pačių autorių surinkti duomenys
Pu et al. (2023) [194]	IEMD-DDCNN	94.49%	Analizuojami tik siurblių duomenys, neskiriant įrenginių ID
Alagele et al. (2024) [195]	AE ir MFCC	93,95%	Taikoma tik dvejetainė normalios ir anomalios būsenos klasifikacija
Chandrakala et al. (2024) [196]	Spektrinė–laikinė sintezė su CLSTM autoenkoderiu	92,94%	Klasifikacija atliekama atskirai kiekvienai įrenginio veikimo būsenai
Zabin et al. (2024) [23]	Savidėmesio SqueezeNet	89,32%	Mokymas ir validacija atliekami tik su –6 dB SNR duomenimis
Zabin et al. (2025) [46]	Few-shot mokymasis su EMD–gammatono spektrograma	89,6%	Mokymas ir validacija atliekami tik su –6 dB SNR duomenimis
Šis darbas	EffNet, SWinT, ir hibridiniai modeliai	99,06%	Mokymas ir validacija atliekami naudojant visą duomenų rinkinį

99,06% tikslumas, patvirtinantis transformerių ir informatyvių laiko ir dažnio požymių derinio efektyvumą. Papildomi rekursiniai sluoksniai SWinT atveju, kaip ir nurodyta [183], reikšmingos naudos nesuteikia.

46 lentelėje matyti, kad ESC-50 duomenų rinkinyje geriausi šiuo metu taikomi metodai pasiekia daugiau nei 97% tikslumą, o dažniausiai tai pasiekiama naudojant išankstinį mokymą arba savimoką [140, 198]. Šiame darbe gautas 93,50% tikslumas yra mažesnis už esamą meno srities lygį, tačiau jis gautas taikant vienodą eksperimentinę procedūrą visiems nagrinėtiems duomenų rinkiniams. Šie rezultatai patvirtina mel-spectrograms pranašumą, palyginti su MFCC ir chroma-STFT, o BiLSTM integravimas į SWinT architektūrą suteikia reikšmingą pranašumą aplinkos garsų klasifikavimui.

Galiausiai, 46 lentelė parodo, kad FSC22 duomenų rinkinyje pasiektas tikslumas yra labiau kintamas dėl jo sudėtingos akustinės struktūros. Siūlomas metodas pranoksta kelis naujausius bazinius modelius, bet nusileidžia geriausiems rezultatams, kuriuos iš dalies lemia skirtingos vertinimo ir duomenų apdorojimo

46 lentelė. Naujausių tyrimų didžiausių pasiekimų palyginimas ESC-50 klasifikavimo uždavinyje

Autorius(-iai)	Metodas	Didžiausias tikslumas
Lin et al. (2020) [137]	ParallelNet	81,55%
Zhou ir Zhao (2022) [138]	TIANNet	84,2%
Li et al. (2022) [139]	Dėmesiu pagrįstas CNN su MFR požymiais	93,1%
Gong et al. (2023) [140]	Garso klasifikavimo metodas, paremtas savimoka ir žinių distiliavimu	97,2%
Liu et al. (2023) [91]	CAT su PANN	96,9%
Sarkar ir Etemad (2022) [197]	CrissCross	79%
Chen et al. (2024) [198]	CL-Transformer	97,75%
Ranmal et al. (2024) [143]	ESC-NAS	81%
Chen et al. (2025) [69]	MobileNetV2 + SPA	91,75%
Šis darbas	EffNet, SWinT, ir hibridiniai modeliai	93,50%

47 lentelė. Naujausių tyrimų didžiausių pasiekimų palyginimas FSC22 klasifikavimo uždavinyje

Autorius(-iai)	Metodas	Didžiausias tikslumas
Bandara et al. (2023) [7]	CNN-pagrįstas modelis su mel-spectrogram	92,59%
Ahmad et al. (2024) [146]	Sluoksniavimo metodas su MFCC	72,7%
Ranmal et al. (2024)	ESC-NAS [143]	85,78%
Simiyu et al. (2024) [199]	Laiko–dažnio CNN	87%
Xu ir Chen (2024) [144]	ERT	66%
Qurthobi et al. (2025) [148]	CNN–BiLSTM su MFCC	78,52%
Sims et al. (2025) [200]	ZeroDiffusion	39,75%
Šis darbas	EffNet, SWinT, ir hibridiniai modeliai	83,16%

schemos. Apskritai mel-spectrograms išlieka tinkamiausia požymių reprezentacija, o rekursinių sluoksnių įtraukimas pabrėžia laikinės informacijos svarbą bioakustinėje klasifikacijoje, nors ir didina skaičiavimo sąnaudas.

Išvados ir ateities darbai

Išvados

1. Pagrindinis šio darbo tikslas buvo sukurti patikimą garso atpažinimo ir klasifikavimo sistemą, galinčią veikti įvairiose aplinkose, ir gauti rezultatai patvirtina, kad šis tikslas buvo pasiektas. Lyginant skirtingus požymių atvaizdavimus (i.e., MFCC, chroma-STFT ir mel-spectrogram) kartu su pažangiomis giliojo mokymosi architektūromis, tokiomis kaip EffNet, SWinT,

ir jų laikiniais variantais, nustatyta, kad siūlomi metodai užtikrina reikšmingą našumo pagerėjimą triukšmingose ir sudėtingose akustinėse aplinkose. Pavyzdžiui, pramoniniame MIMII duomenų rinkinyje siūlomi hibridiniai modeliai pasiekė didesnį nei $99,0\%$ klasifikavimo tikslumą, viršydami ankstesnius darbus, kuriuose buvo nurodomas maždaug $89,6\%$ tikslumas. Atitinkamai, ESC-50 duomenų rinkinyje modeliai, integruojantys SWinT su laikiniais sluoksniais, pasiekė didesnį nei $72,3\% \pm 2,4\%$ tikslumą, daugiau nei 10 procentinių punktų viršydami tradicinius dėmesio mechanizmais pagrįstus transformerius. Gamtinėje aplinkoje FSC22 duomenų rinkinyje mel-spectrogram ir SWinT-BiLSTM deriniai pasiekė iki $82,5\% \pm 1,0\%$ tikslumą, parodant sistemos gebėjimą apdoroti persidengiančius natūralius garsus. Šie rezultatai patvirtina, kad pasiūlyta sistema veiksmingai sprendžia domenų įvairovės ir foninio triukšmo problemas. Nors eksperimentai atlikti taikant prižiūrimos klasifikacijos metodus, gauti rezultatai tiesiogiai pagrindžia anomalijų aptikimą, leidžiant patikimai atskirti normalius ir nenormalius akustinius modelius triukšmingose aplinkose.

2. MIMII duomenų rinkinys, palyginti su ESC-50 ir FSC22, pasižymi ryškiu disbalansu tarp normalių ir anomalinių įrašų, tiek tarp skirtingų įrenginių tipų, tiek tarp to paties įrenginio identifikatorių. Šis disbalansas trukdo patikimai aptikti anomalijas. Siekiant sumažinti šią problemą, darbe taikyta SNR-pagrindu paremta duomenų augmentacija, naudojant esamus įrašus su skirtingais triukšmo lygiais (-6 dB, 0 dB ir 6 dB). Šis metodas išplėtė efektyvią mokymo imtį ir įvedė papildomą akustinę įvairovę, leidžiančią pasiekti iki $99,06\% \pm 0,02\%$ tikslumą, viršijant anksčiau skelbtus rezultatus. Pagerėjo visi nagrinėti požymiai, o didžiausias efektas pasireiškė triukšmingomis sąlygomis, kada anksčiau buvo fiksuojamas persimokymas.
3. Trijų pagrindinių garso požymių (mel-spectrogram, MFCC ir chroma-STFT) lyginamoji analizė atskleidė skirtingas našumo tendencijas, priklausančias nuo jų sąveikos su modelių architektūromis ir duomenų rinkinių savybėmis. mel-spectrogram, užtikrinanti detalią laiko ir dažnio raišką, nuosekliai demonstruoja aukštą klasifikavimo našumą, ypač FSC22 duomenų rinkinyje derinyje su SWinT-BiLSTM, kur buvo pasiektas $82,47\% \pm 0,98\%$ F1 rodiklis. MFCC pasižymi palankiu skaičiavimo sudėtingumo ir tikslumo balansu, užtikrindama stabilų veikimą ESC-50 rinkinyje. O chroma-STFT, pabrėžianti toninę informaciją, pasirodė silpniausiai netoninių garso klasifikavimo užduočių atveju, ypač MIMII ir FSC22 rinkiniuose. Šias išvadas papildomai patvirtina t-SNE vizualizacijos, atskleidžiančios prastesnį klasių atskyrimą naudojant chroma-STFT požymius.
4. Siekiant pagerinti klasifikavimo rezultatus, darbe taikytas TL iš anksto išmokytiems EffNet ir SWinT modeliams, kurie buvo išplėsti laikiniais moduliais, tokiais kaip BiGRU ir BiLSTM. Gauti hibridiniai modeliai (EffNet-BiGRU, EffNet-BiLSTM, SWinT-BiGRU, SWinT-BiLSTM) efektyviai fiksuoja laiko dinamiką, ypač duomenų rinkiniuose su greitai kintančiais ar persidengiančiais garso šaltiniais.

Reikšmingi pagerėjimai pasiekti ESC-50 ir FSC22 duomenų rinkiniuose, patvirtinant, kad laikinis modeliavimas yra ypač naudingas sudėtingose akustinėse užduotyse.

5. Siūloma garso klasifikavimo sistema buvo įvertinta trijuose duomenų rinkiniuose (MIMII, ESC-50 ir FSC22), atspindinčiuose pramonines, miesto ir gamtines aplinkas. Vertinimas atliktas naudojant tikslumą, AUC, Preciziškumo, atkuriamumą (*recall*) ir F1 rodiklį. Hibridiniai modeliai, ypač SWinT-BiLSTM ir SWinT-BiGRU, nuosekliai pasiekė aukštą našumą visuose rinkiniuose, o t-SNE analizė parodė geresnį klasių atskyrimą naudojant MFCC ir mel-spectrogram požymius.
6. Eksperimentai taip pat suteikė įžvalgų apie mokymo dinamiką, ypač ankstyvo stabdymo taikymą, kai validavimo nuostoliai negerėjo nustatytą epochų skaičių. Nė vienas modelis nepasiekė maksimalaus epochų skaičiaus, o tai rodo, kad ankstyvas stabdymas veiksmingai riboja persimokymą. Šie rezultatai leidžia daryti išvadą, kad rekursiniai sluoksniai prisideda prie stabilesnio mokymo ir geresnio konvergavimo, ypač derinant su konvoliucinėmis architektūromis.

Ateities darbai

1. Ateityje DL bus taikomas neįprastose aplinkose, tokiose kaip po vandeniu, kur dėl specifinių sąlygų sklinda kitokie garsai, pvz., jūrų žinduolių balsai, sonarų signalai, laivų keliamas triukšmas ar geologiniai garsai. Tokios aplinkybės reikalauja unikalių spektrinių reprezentacijų ir hibridinių modelių, įgalinančių dėmesio mechanizmus ir pasikartojančius sluoksnius, kad jūrų ekosistemas būtų galima stebėti realiu laiku ir užtikrinti saugumą [49].
2. Siekiamybė yra sukurti garso rinkinius, kurie yra specialiai pritaikyti konkretiems regionams ir atspindi skirtingą klimatą, biologinę įvairovę, kultūrą bei kalbą. Šie duomenys padidintų modelių tikslumą atliekant regionines užduotis, tokias kaip laukinės gamtos stebėjimas ar miesto triukšmo analizė, taip pat leistų vykdyti geoakustines studijas. Tokių rinkinių rinkimui ir naudojimui būtinas institucijų ir bendruomenių bendradarbiavimas.

REFERENCES

1. KILICKAYA, S., AHISHALI, M., CELEBIOGLU, C., SOHRAB, F., EREN, L., INCE, T., ASKAR, M., and GABBOUJ, M. Audio-based Anomaly Detection in Industrial Machines Using Deep One-Class Support Vector Data Description. In: *2025 IEEE Symposium on Computational Intelligence on Engineering/Cyber Physical Systems Companion (CIES Companion)*. IEEE, 2025, pp. 1–5. Available from DOI: 10.1109/ciescompanion65073.2025.11010815.

2. LO SCUDO, F., RITACCO, E., CAROPRESE, L., and MANCO, G. Audio-based anomaly detection on edge devices via self-supervision and spectral analysis. *Journal of Intelligent Information Systems*. 2023, vol. 61, pp. 1–29. Available from DOI: 10.1007/s10844-023-00792-2.

3. BARUSCO, M., BORSATTI, F., PEZZE, D. D., PAISSAN, F., FARELLA, E., and SUSTO, G. A. *From Vision to Sound: Advancing Audio Anomaly Detection with Vision-Based Algorithms*. 2025. Available from arXiv: 2502.18328 [cs.SD].

4. PUROHIT, H., TANABE, R., ICHIGE, K., ENDO, T., NIKAIDO, Y., SUEFUSA, K., and KAWAGUCHI, Y. *MIMII Dataset: Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection*. 2019. Available from arXiv: 1909.09347 [cs.SD].

5. PICZAK, K. J. ESC: Dataset for Environmental Sound Classification. In: *Proceedings of the 23rd ACM International Conference on Multimedia*. Brisbane, Australia: Association for Computing Machinery, 2015, pp. 1015–1018. MM '15. ISBN 9781450334594. Available from DOI: 10.1145/2733373.2806390.

6. SALAMON, J., JACOBY, C., and BELLO, J. P. A Dataset and Taxonomy for Urban Sound Research. In: *Proceedings of the 22nd ACM International Conference on Multimedia*. Orlando, Florida, USA: Association for Computing Machinery, 2014, pp. 1041–1044. MM '14. ISBN 9781450330633. Available from DOI: 10.1145/2647868.2655045.

7. BANDARA, M., JAYASUNDARA, R., ARIYARATHNE, I., MEEDENIYA, D., and PERERA, C. Forest Sound Classification Dataset: FSC22. *Sensors*. 2023, vol. 23, pp. 1–22. Available from DOI: 10.3390/s23042032.

8. KHANJARI, M., AZARFAR, A., HOSSEINI ABARDEH, M., and ALIBEIKI, E. Anomalous sound detection for machine condition monitoring using 3D tensor representation of sound and 3D deep convolutional neural network. *Multimedia Tools and Applications*. 2023, vol. 83, pp. 1–19. Available from DOI: 10.1007/s11042-023-17043-9.

9. KIM, E., MUN, D., JUN, M., and YUN, H. Operation and Productivity Monitoring from Sound Signal of Legacy Pipe Bending Machine via Convolutional Neural Network (CNN). *International Journal of Precision Engineering and Manufacturing*. 2024, vol. 25. Available from DOI: 10.1007/s12541-024-01018-3.

10. WEGENER, K., BLEICHER, F., HEISEL, U., HOFFMEISTER, H.-W., and MÖHRING, H.-C. Noise and vibrations in machine tools. *CIRP Annals*. 2021, vol. 70, no. 2, pp. 611–633. ISSN 0007-8506. Available from DOI: https://doi.org/10.1016/j.cirp.2021.05.010.

11. GUAN, J., LIU, Y., KONG, Q., XIAO, F., ZHU, Q., TIAN, J., and WANG, W. *Transformer-based Autoencoder with ID Constraint for Unsupervised Anomalous Sound Detection*. 2023. Available from arXiv: 2310.08950 [cs.SD].

12. JOMBO, G., and ZHANG, Y. Acoustic-Based Machine Condition Monitoring—Methods and Challenges. *Eng*. 2023, vol. 4, no. 1, pp. 47–79. ISSN 2673-4117. Available from DOI: 10.3390/eng4010004.

13. TANEJA, J., KRIOUKOV, A., DAWSON-HAGGERTY, S., and CULLER, D. Enabling advanced environmental conditioning with a building application stack. In: *2013 International Green Computing Conference Proceedings*. 2013, pp. 1–10. Available from DOI: 10.1109/IGCC.2013.6604519.

14. MONTES GONZÁLEZ, D., BARRIGÓN MORILLAS, J. M., and REY-GOZALO, G. Effects of noise on pedestrians in urban environments where road traffic is the main source of sound. *Science of The Total Environment*. 2023, vol. 857, p. 159406. ISSN 0048-9697. Available from DOI: https://doi.org/10.1016/j.scitotenv.2022.159406.

15. PAZ, E. C. da, VIEIRA, T. J., and ZANNIN, P. H. T. Urban Noise as an Environmental Impact Factor in the Urban Planning Process. In: ERGEN, Y. B. (ed.). *An Overview of Urban and Regional Planning*. Rijeka: IntechOpen, 2018, chap. 2. Available from DOI: 10.5772/intechopen.78643.

16. TSUKERNIKOV, I., ANTONOV, A., LEDENEV, V., SHUBIN, I., and NEVENCHANNAYA, T. Acoustic Characteristics Analysis of Industrial Premises with Process Equipment. *Journal of Applied Mathematics and Physics*. 2016, vol. 04, pp. 206–210. Available from DOI: https://doi.org/10.4236/jamp.2016.42026.

17. GEMMEKE, J. F., ELLIS, D. P. W., FREEDMAN, D., JANSEN, A., LAWRENCE, W., MOORE, R. C., PLAKAL, M., and RITTER, M. 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. Available from DOI: 10.1109/ICASSP.2017.7952261.

18. SNYDER, D., GARCIA-ROMERO, D., SELL, G., POVEY, D., and KHUDANPUR, S. X-Vectors: Robust DNN Embeddings for Speaker Recognition. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018, pp. 5329–5333. Available from DOI: 10.1109/ICASSP.2018.8461375.

19. BARCHIESI, D., GIANNOULIS, D., STOWELL, D., and PLUMBLEY, M. D. Acoustic Scene Classification: Classifying environments from the sounds they produce. *IEEE Signal Processing Magazine*. 2015, vol. 32, no. 3, pp. 16–34. Available from DOI: 10.1109/MSP.2014.2326181.

20. BADAJOS, J. C. T., OCHOA, K. M. P., TEJADILLA, R. A. P., and PEÑAS, R. T. L. Reduction of Audio Noise with Lowpass Chebyshev Type II Filter Simulated using GNU Octave. In: *2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*. 2019, pp. 1–6. Available from DOI: 10.1109/HNICEM48295.2019.9072844.

21. AHN, H., and YEO, I. Deep-Learning-Based Approach to Anomaly Detection Techniques for Large Acoustic Data in Machine Operation. *Sensors*. 2021, vol. 21, no. 16. ISSN 1424-8220. Available from DOI: 10.3390/s21165446.

22. NISHIDA, T., HARADA, N., NIIZUMI, D., ALBERTINI, D., SANNINO, R., PRADOLINI, S., AUGUSTI, F., IMOTO, K., DOHI, K., PUROHIT, H., ENDO, T., and KAWAGUCHI, Y. *Description and Discussion on DCASE 2024 Challenge Task 2: First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring*. 2024. Available from arXiv: 2406.07250 [eess.AS].

23. ZABIN, M., BINTE KABIR, A. N., KABIR, M. K., CHOI, H.-J., and UDDIN, J. Machine Fault Diagnosis Using EMD-Gammatone Texture Representation and A Lightweight Self-Attention SqueezeNet. In: *2024 IEEE International Conference on Big Data and Smart Computing (BigComp)*. 2024, pp. 32–39. Available from DOI: 10.1109/BigComp60711.2024.00015.

24. ZAHEER, R., AHMAD, I., HABIBI, D., ISLAM, K. Y., and PHUNG, Q. V. A Survey on Artificial Intelligence-Based Acoustic Source Identification. *IEEE Access*. 2023, vol. 11, pp. 60078–60108. Available from DOI: 10.1109/ACCESS.2023.3283982.

25. NTALAMPIRAS, S. One-shot learning for acoustic diagnosis of industrial machines. *Expert Systems with Applications*. 2021, vol. 178, p. 114984. ISSN 0957-4174. Available from DOI: https://doi.org/10.1016/j.eswa.2021.114984.

26. KANG, D.-J., GU, J.-H., and LEE, J.-W. Characteristics of Industrial Machinery Noise. *Transactions of The Korean Society for Noise and Vibration Engineering*. 2010, vol. 20, pp. 160–165. Available from DOI: https://doi.org/10.5050/KSNVE.2010.20.2.160.

27. SHAIKH, K. B. T., JAWARKAR, N. P., and AHMED, V. Machine diagnosis using acoustic analysis: a review. In: *2021 IEEE Conference on Norbert Wiener in the 21st Century (21CW)*. 2021, pp. 1–6. Available from DOI: 10.1109/21CW48944.2021.9532537.

28. BENOCCI, R., ROMAN, H. E., BISCEGLIE, A., ANGELINI, F., BRAMBILLA, G., and ZAMBON, G. Eco-Acoustic Assessment of an Urban Park by Statistical Analysis. *Sustainability*. 2021, vol. 13, no. 14, p. 7857. ISSN 2071-1050. Available from DOI: 10.3390/su13147857.

29. JIN, W., WANG, X., and ZHAN, Y. Environmental Sound Classification Algorithm Based on Region Joint Signal Analysis Feature and Boosting Ensemble Learning. *Electronics*. 2022, vol. 11, p. 3743. Available from DOI: https://doi.org/10.3390/electronics11223743.

30. MEEDENIYA, D., ARIYARATHNE, I., BANDARA, M., JAYASUNDARA, R., and PERERA, C. A Survey on Deep Learning Based Forest Environment Sound Classification at the Edge. *ACM Comput. Surv.* 2023, vol. 56, no. 3. ISSN 0360-0300. Available from DOI: 10.1145/3618104.

31. SEGARCEANU, S., OLTEANU, E., and SUCIU, G. Forest Monitoring Using Forest Sound Identification. In: *2020 43rd International Conference on Telecommunications and Signal Processing (TSP)*. 2020, pp. 346–349. Available from DOI: 10.1109/TSP49548.2020.9163433.

32. OLTEANU, E., SUCIU, V., SEGARCEANU, S., PETRE, I., and SCHEIANU, A. Forest Monitoring System Through Sound Recognition. In: *2018 International Conference on Communications (COMM)*. 2018, pp. 75–80. Available from DOI: 10.1109/ICComm.2018.8484773.

33. DAMASEVICIUS, R., QURTHOBI, A., and MASKELIUNAS, R. A Hybrid Machine Learning Model for Forest Wildfire Detection using Sounds. In: *2024 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*. 2024, pp. 99–106. Available from DOI: 10.15439/2024F7263.

34. PANDYA, S., and GHAYVAT, H. Ambient acoustic event assistive framework for identification, detection, and recognition of unknown acoustic events of a residence. *Advanced Engineering Informatics*. 2021, vol. 47, p. 101238. ISSN 1474-0346. Available from DOI: 10.1016/j.aei.2020.101238.

35. CHAKMA, A., DAS, A., MD FARIDEE, A. Z., CHAKRABORTY, S., CHAKRABORTY, S., and ROY, N. AcouDL: Context-Aware Daily Activity Recognition from Natural Acoustic Signals. In: *2024 IEEE International Conference on Smart Computing (SMARTCOMP)*. 2024, pp. 332–337. Available from DOI: 10.1109/SMARTCOMP61445.2024.00077.

36. MASKELIUNAS, R., RAUDONIS, V., and DAMASEVICIUS, R. Recognition of Emotional Vocalizations of Canine. *Acta Acustica united with Acustica*. 2018, vol. 104, no. 2, pp. 304–314. ISSN 1610-1928. Available from DOI: 10.3813/aaa.919173.

37. FONT, F., ROMA, G., and SERRA, X. Freesound technical demo. In: *Proceedings of the 21st ACM International Conference on Multimedia*. Barcelona, Spain: Association for Computing Machinery, 2013, pp. 411–412. MM '13. ISBN 9781450324045. Available from DOI: 10.1145/2502081.2502245.

38. MESAROS, A., HEITTOLA, T., and VIRTANEN, T. *A multi-device dataset for urban acoustic scene classification*. 2018. Available from arXiv: 1807.09840 [eess.AS].

39. KOIZUMI, Y., SAITO, S., UEMATSU, H., HARADA, N., and IMOTO, K. *ToyADMOS: A Dataset of Miniature-Machine Operating Sounds for Anomalous Sound Detection*. 2019. Available from arXiv: 1908.03299 [eess.AS].

40. FONSECA, E., FAVORY, X., PONS, J., FONT, F., and SERRA, X. *FSD50K: An Open Dataset of Human-Labeled Sound Events*. 2022. Available from arXiv: 2010.00475 [cs.SD].

41. STEVENS, S. S., VOLKMANN, J., and NEWMAN, E. B. A Scale for the Measurement of the Psychological Magnitude Pitch. *The Journal of the Acoustical Society of America*. 1937, vol. 8, no. 3, pp. 185–190. ISSN 0001-4966. Available from DOI: 10.1121/1.1915893.

42. DAVIS, S., and MERMELSTEIN, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*. 1980, vol. 28, no. 4, pp. 357–366. Available from DOI: 10.1109/TASSP.1980.1163420.

43. ELLIS, D. P. W. Beat Tracking by Dynamic Programming. *Journal of New Music Research*. 2007, vol. 36, no. 1, pp. 51–60. Available from DOI: 10.1080/09298210701653344.

44. ELLIS, D. P., and POLINER, G. E. Identifying 'Cover Songs' with Chroma Features and Dynamic Programming Beat Tracking. In: *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*. 2007, vol. 4, pp. IV-1429-IV–1432. Available from DOI: 10.1109/ICASSP.2007.367348.

45. MÜLLER, M., and EWERT, S. Chroma Toolbox: Matlab Implementations for Extracting Variants of Chroma-Based Audio Features. In: KLAPURI, A., and LEIDER, C. (eds.). *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, Miami, Florida, USA, October 24-28, 2011*. University of Miami, 2011, pp. 215–220. Available also from: http://ismir2011.ismir.net/papers/PS2-8.pdf.

46. ZABIN, M., AYON, S. T. K., SIRAJ, F. M., SHUVO, M. H., CHOI, H.-J., and UDDIN, J. Few-Shot Learning-Based Machine Fault Diagnosis Using EMD-Gammatone Spectrogram with Limited Labeled Audio Dataset. In: *2025 IEEE International Conference on Big Data and Smart Computing (BigComp)*. 2025, pp. 183–190. Available from DOI: 10.1109/BigComp64353.2025.00046.

47. COLLACOTT, R. Condition monitoring by sound analysis. *Non-Destructive Testing*. 1975, vol. 8, no. 5, pp. 245–248. ISSN 0029-1021. Available from DOI: https://doi.org/10.1016/0029-1021(75)90044-4.

48. TAGAWA, Y., MASKELIŪNAS, R., and DAMAŠEVIČIUS, R. Acoustic Anomaly Detection of Mechanical Failures in Noisy Real-Life Factory Environments. *Electronics*. 2021, vol. 10, no. 19. ISSN 2079-9292. Available from DOI: 10.3390/electronics10192329.

49. TSUJI, K., IMAI, S., TAKAO, R., KIMURA, T., KONDO, H., and AND, Y. K. A machine sound monitoring for predictive maintenance focusing on very low frequency band. *SICE Journal of Control, Measurement, and System Integration*. 2021, vol. 14, no. 1, pp. 27–38. Available from DOI: 10.1080/18824889.2020.1863611.

50. OTA, Y., and UNOKI, M. Anomalous Sound Detection for Industrial Machines Using Acoustical Features Related to Timbral Metrics. *IEEE Access*. 2023, vol. 11, pp. 70884–70897. Available from DOI: 10.1109/ACCESS.2023.3294334.

51. LEE, Y., KIM, J., and OK, J. *Activity-Guided Industrial Anomalous Sound Detection against Interferences*. 2024. Available from arXiv: 2409.01885 [cs.SD].

52. TANG, L., TIAN, H., HUANG, H., SHI, S., and JI, Q. A survey of mechanical fault diagnosis based on audio signal analysis. *Measurement*. 2023, vol. 220, p. 113294. ISSN 0263-2241. Available from DOI: https://doi.org/10.1016/j.measurement.2023.113294.

53. STOWELL, D. Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*. 2022, vol. 10, e13152. ISSN 2167-8359. Available from DOI: 10.7717/peerj.13152.

54. XIE, J., ZHONG, Y., ZHANG, J., LIU, S., DING, C., and TRIANTAFYLLOPOULOS, A. A review of automatic recognition technology for bird vocalizations in the deep learning era. *Ecological Informatics*. 2023, vol. 73, p. 101927. ISSN 1574-9541. Available from DOI: https://doi.org/10.1016/j.ecoinf.2022.101927.

55. AYERS, J. G., JANDALI, Y., HWANG, Y.-J., JOUN, E., STEINBERG, G., TOBLER, M., INGRAM, I., KASTNER, R., and SCHURGERS, C. Challenges in Applying Audio Classification Models to Datasets Containing Crucial Biodiversity Information. In: *ICML 2021 Workshop on Tackling Climate Change with Machine Learning*. 2021. Available also from: https://www.climatechange.ai/papers/icml2021/14.

56. ALVAREZ, A. A., and GÓMEZ, F. Motivic Pattern Classification of Music Audio Signals Combining Residual and LSTM Networks. *International Journal of Interactive Multimedia and Artificial Intelligence*. 2021, vol. 6, no. 6, pp. 208–214. ISSN 1989-1660. Available from DOI: 10.9781/ijimai.2021.01.003.

57. MOBTAHEJ, P., ZHANG, X., HAMIDI, M., and ZHANG, J. An LSTM-Autoencoder Architecture for Anomaly Detection Applied on Compressors Audio Data. *Computational and Mathematical Methods*. 2022, vol. 2022, no. 1, p. 3622426. Available from DOI: https://doi.org/10.1155/2022/3622426.

58. DUAN, G., SU, Y., and FU, J. Landslide Displacement Prediction Based on Multivariate LSTM Model. *International Journal of Environmental Research and Public Health*. 2023, vol. 20, p. 1167. Available from DOI: 10 . 3390 / ijerph20021167.

59. SAMS, A., and ZAHRA, A. Multimodal music emotion recognition in Indonesian songs based on CNN-LSTM, XLNet transformers. *Bulletin of Electrical Engineering and Informatics*. 2023, vol. 12, pp. 355–364. Available from DOI: https://doi.org/10.11591/eei.v12i1.4231.

60. AQEEL, A., HASSAN, A., KHAN, M., REHMAN, S., TARIQ, U., KADRY, S., MAJUMDAR, A., and THINNUKOOL, O. A Long Short-Term Memory Biomarker-Based Prediction Framework for Alzheimer's Disease. *Sensors*. 2022, vol. 22, pp. 1–15. Available from DOI: 10.3390/s22041475.

61. KASTHURI, E., and BALAJI, S. Natural language processing and deep learning chatbot using long short term memory algorithm. *Materials Today: Proceedings*. 2021, vol. 81, pp. 690–693. Available from DOI: https://doi.org/10.1016/j.matpr.2021.04.154.

62. FJELLSTRÖM, C. Long Short-Term Memory Neural Network for Financial Time Series. In: *2022 IEEE International Conference on Big Data (Big Data)*. 2022, pp. 3496–3504. Available from DOI: 10.1109/BigData55660.2022.10020784.

63. CHO, K., MERRIENBOER, B. van, GULCEHRE, C., BAHDANAU, D., BOUGARES, F., SCHWENK, H., and BENGIO, Y. *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*. 2014. Available from arXiv: 1406.1078 [cs.CL].

64. LIU, T., JIN, Y., WANG, S., ZHENG, Q., and YANG, G. Denoising method of weak fault acoustic emission signal under strong background noise of engine based on autoencoder and wavelet packet decomposition. *Structural Health Monitoring*. 2023, vol. 22, no. 5, pp. 3206–3224. Available from DOI: 10.1177/14759217221143547.

65. TRAN, T., BADER, S., and LUNDGREN, J. Denoising Induction Motor Sounds Using an Autoencoder. In: *2023 IEEE Sensors Applications Symposium (SAS)*. 2023, pp. 01–06. Available from DOI: 10.1109/SAS58821.2023.10254150.

66. ZHAO, Y., HAO, H., CHEN, Y., and ZHANG, Y. Novelty Detection and Fault Diagnosis Method for Bearing Faults Based on the Hybrid Deep Autoencoder Network. *Electronics*. 2023, vol. 12, p. 2826. Available from DOI: 10 . 3390 / electronics12132826.

67. PUROHIT, H., ENDO, T., YAMAMOTO, M., and KAWAGUCHI, Y. *Hierarchical Conditional Variational Autoencoder Based Acoustic Anomaly Detection*. 2022. Available from arXiv: 2206.05460 [cs.LG].

68. PENG, B., LI, D., WANG, K., and ABDULLA, W. Acoustic-Based Industrial Diagnostics: A Scalable Noise-Robust Multiclass Framework for Anomaly Detection. *Processes*. 2025, vol. 13, p. 544. Available from DOI: 10.3390/pr13020544.

69. CHEN, F., ZHU, Z., SUN, C., and XIA, L. Evaluating metric and contrastive learning in pretrained models for environmental sound classification. *Applied Acoustics*. 2025, vol. 232, p. 110593. Available from DOI: 10.1016/j.apacoust.2025.110593.

70. HUANG, G., LIU, Z., MAATEN, L. van der, and WEINBERGER, K. Q. *Densely Connected Convolutional Networks*. 2018. Available from arXiv: 1608 . 06993 [cs.CV].

71. TAN, M., and LE, Q. V. *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. 2020. Available from arXiv: 1905.11946 [cs.LG].

72. TAN, M., and LE, Q. V. *EfficientNetV2: Smaller Models and Faster Training*. 2021. Available from arXiv: 2104.00298 [cs.CV].

73. SANDLER, M., HOWARD, A., ZHU, M., ZHMOGINOV, A., and CHEN, L.-C. *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. 2019. Available from arXiv: 1801.04381 [cs.CV].

74. HOWARD, A., SANDLER, M., CHU, G., CHEN, L.-C., CHEN, B., TAN, M., WANG, W., ZHU, Y., PANG, R., VASUDEVAN, V., LE, Q. V., and ADAM, H. *Searching for MobileNetV3*. 2019. Available from arXiv: 1905.02244 [cs.CV].

75. LIU, Z., MAO, H., WU, C.-Y., FEICHTENHOFER, C., DARRELL, T., and XIE, S. *A ConvNet for the 2020s*. 2022. Available from arXiv: 2201.03545 [cs.CV].

76. LIU, Z., LIN, Y., CAO, Y., HU, H., WEI, Y., ZHANG, Z., LIN, S., and GUO, B. *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. 2021. Available from arXiv: 2103.14030 [cs.CV].

77. LIU, Z., HU, H., LIN, Y., YAO, Z., XIE, Z., WEI, Y., NING, J., CAO, Y., ZHANG, Z., DONG, L., WEI, F., and GUO, B. *Swin Transformer V2: Scaling Up Capacity and Resolution*. 2022. Available from arXiv: 2111.09883 [cs.CV].

78. DOSOVITSKIY, A., BEYER, L., KOLESNIKOV, A., WEISSENBORN, D., ZHAI, X., UNTERTHINER, T., DEHGHANI, M., MINDERER, M., HEIGOLD, G., GELLY, S., USZKOREIT, J., and HOULSBY, N. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. Available from arXiv: 2010.11929 [cs.CV].

79. ZADA, S., BENOU, I., and IRANI, M. *Pure Noise to the Rescue of Insufficient Data: Improving Imbalanced Classification by Training on Random Noise Images*. 2022. Available from arXiv: 2112.08810 [cs.CV].

80. JAISWAL, M., and PROVOST, E. M. *Best Practices for Noise-Based Augmentation to Improve the Performance of Deployable Speech-Based Emotion Recognition Systems*. 2023. Available from arXiv: 2104.08806 [cs.SD].

81. SEJDIĆ, E., DJUROVIĆ, I., and JIANG, J. Time–frequency feature representation using energy concentration: An overview of recent advances. *Digital Signal Processing*. 2009, vol. 19, no. 1, pp. 153–183. ISSN 1051-2004. Available from DOI: https://doi.org/10.1016/j.dsp.2007.12.004.

82. ABDUL, Z. K., and AL-TALABANI, A. K. Mel Frequency Cepstral Coefficient and its Applications: A Review. *IEEE Access*. 2022, vol. 10, pp. 122136–122158. Available from DOI: 10.1109/ACCESS.2022.3223444.

83. MÜLLER, M. *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Cham: Springer, 2015. Available from DOI: 10.1007/978-3-319-21945-5.

84. HUMPHREY, E. J., BELLO, J. P., and LECUN, Y. Moving Beyond Feature Design: Deep Architectures and Automatic Feature Learning in Music Informatics. In: *Proceedings of the 13th International Society for Music Information Retrieval Conference*. ISMIR, 2012, pp. 403–408. Available from DOI: 10.5281/zenodo.1415726.

85. DIELEMAN, S., and SCHRAUWEN, B. End-to-end learning for music audio. In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2014, pp. 6964–6968. Available from DOI: 10.1109/ICASSP.2014.6854950.

86. KONG, Q., CAO, Y., IQBAL, T., WANG, Y., WANG, W., and PLUMBLEY, M. D. *PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition*. 2020. Available from arXiv: 1912.10211 [cs.SD].

87. SALAMON, J., and BELLO, J. P. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. *IEEE Signal Processing Letters*. 2017, vol. 24, no. 3, pp. 279–283. ISSN 1558-2361. Available from DOI: 10.1109/lsp.2017.2657381.

88. GONG, Y., CHUNG, Y.-A., and GLASS, J. PSLA: Improving Audio Tagging With Pretraining, Sampling, Labeling, and Aggregation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2021, vol. 29, pp. 3292–3306. Available from DOI: 10.1109/TASLP.2021.3120633.

89. PORWAL, A. Bird-Species Audio Identification, Ensembling of EfficientNet-B0 and Pre-trained EfficientNet-B1 model. In: *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (CEUR-WS)*. 2024, vol. 3740. ISSN 1613-0073. Available also from: https://ceur-ws.org/Vol-3740/paper-204.pdf.

90. WANG, M., and YANG, Z. TFECN: Time-Frequency Enhanced ConvNet for Audio Classification. In: *Proc. Interspeech 2023*. 2023, pp. 281–285. Available from DOI: 10.21437/Interspeech.2023-734.

91. LIU, X., LU, H., YUAN, J., and LI, X. *CAT: Causal Audio Transformer for Audio Classification*. 2023. Available from arXiv: 2303.07626 [cs.SD].

92. ZHAO, H., ZHANG, C., ZHU, B., MA, Z., and ZHANG, K. *S3T: Self-Supervised Pre-training with Swin Transformer for Music Classification*. 2022. Available from arXiv: 2202.10139 [eess.AS].

93. TSALERA, E., PAPADAKIS, A., and SAMARAKOU, M. Comparison of Pre-Trained CNNs for Audio Classification Using Transfer Learning. *Journal of Sensor and Actuator Networks*. 2021, vol. 10, no. 4. ISSN 2224-2708. Available from DOI: 10.3390/jsan10040072.

94. SIMONYAN, K., and ZISSERMAN, A. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. Available from arXiv: 1409.1556 [cs.CV].

95. HERSHEY, S., CHAUDHURI, S., ELLIS, D. P. W., GEMMEKE, J. F., JANSEN, A., MOORE, R. C., PLAKAL, M., PLATT, D., SAUROUS, R. A., SEYBOLD, B., SLANEY, M., WEISS, R. J., and WILSON, K. *CNN Architectures for Large-Scale Audio Classification*. 2017. Available from arXiv: 1609.09430 [cs.SD].

96. CHOUDHARY, S., KARTHIK, C. R., LAKSHMI, P. S., and KUMAR, S. LEAN: Light and Efficient Audio Classification Network. In: *2022 IEEE 19th India Council International Conference (INDICON)*. 2022, pp. 1–6. Available from DOI: 10.1109/INDICON56171.2022.10039921.

97. HE, H., and LUO, H. An improved lightweight method based on EfficientNet for birdsong recognition. *Scientific Reports*. 2025, vol. 15, no. 1, p. 23727. Available from DOI: 10.1038/s41598-025-07875-w.

98. TZANETAKIS, G., and COOK, P. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*. 2002, vol. 10, no. 5, pp. 293–302. Available from DOI: 10.1109/TSA.2002.800560.

99. STURM, B. L. The State of the Art Ten Years After a State of the Art: Future Research in Music Information Retrieval. *Journal of New Music Research*. 2014, vol. 43, no. 2, pp. 147–172. ISSN 1744-5027. Available from DOI: 10.1080/09298215.2014.894533.

100. DUAN, Y. Broadcast Swin Transformer for Music Genre Classification. *Highlights in Science, Engineering and Technology*. 2024, vol. 85, pp. 691–699. Available from DOI: https://doi.org/10.54097/p1x81q26.

101. WANG, Y., LU, C., LIAN, H., ZHAO, Y., SCHULLER, B. W., ZONG, Y., and ZHENG, W. Speech Swin-Transformer: Exploring a Hierarchical Transformer with Shifted Windows for Speech Emotion Recognition. In: *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2024, pp. 11646–11650. Available from DOI: 10.1109/ICASSP48485.2024.10447726.

102. RAMESH, R., PRAHALADHAN, V., NITHISH, P., and KOTHANDARAMAN, M. Speech emotion recognition using the novel SwinEmoNet (Shifted Window Transformer Emotion Network). *International Journal of Speech Technology*. 2024, vol. 27, pp. 551–568. Available from DOI: 10.1007/s10772-024-10123-7.

103. BURKHARDT, F., PAESCHKE, A., KIENAST, M., SENDLMEIER, W. F., and WEISS, B. *Berlin EmoDB*. Zenodo, 2022. Version 1.3.0. Available from DOI: 10.5281/zenodo.7447302.

104. LIVINGSTONE, S. R., and RUSSO, F. A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one*. 2018, vol. 13, no. 5, e0196391. Available from DOI: 10.1371/journal.pone.0196391.

105. SIMONOVIĆ, M., KOVANDŽIĆ, M., ĆIRIĆ, I., and NIKOLIĆ, V. Acoustic recognition of noise-like environmental sounds by using artificial neural network. *Expert Systems with Applications*. 2021, vol. 184, p. 115484. ISSN 0957-4174. Available from DOI: https://doi.org/10.1016/j.eswa.2021.115484.

106. MOLINA VICUNA, C., and HOWELER, C. A method for reduction of Acoustic Emission (AE) data with application in machine failure detection and diagnosis. *Mechanical Systems and Signal Processing*. 2017, vol. 97. Available from DOI: 10.1016/j.ymssp.2017.04.040.

107. HOLFORD, K., EATON, M., HENSMAN, J., PULLIN, R., EVANS, S., DERVILIS, N., and WORDEN, K. A new methodology for automating acoustic emission detection of metallic fatigue fractures in highly demanding aerospace environments: An overview. *Progress in Aerospace Sciences*. 2017, vol. 90. Available from DOI: 10.1016/j.paerosci.2016.11.003.

108. COOPER, C., WANG, P., ZHANG, J., GAO, R., RONEY, T., RAGAI, I., and SHAFFER, D. Convolutional neural network-based tool condition monitoring in vertical milling operations using acoustic signals. *Procedia Manufacturing*. 2020, vol. 49, pp. 105–111. Available from DOI: 10.1016/j.promfg.2020.07.004.

109. REUBENS, J., VERHELST, P., KNAAP, I. van der, DENEUDT, K., MOENS, T., and HERNANDEZ, F. Environmental factors influence the detection probability in acoustic telemetry in a marine environment: results from a new setup. *Hydrobiologia*. 2019, vol. 845, pp. 81–94. Available from DOI: 10.1007/s10750-017-3478-7.

110. FAN, L., ZHANG, F., FAN, H., and ZHANG, C. Brief review of image denoising techniques. *Visual Computing for Industry, Biomedicine, and Art*. 2019, vol. 2, no. 1. Available from DOI: 10.1186/s42492-019-0016-7.

111. ALI, M. H., JABER, M. M., ABD, S. K., REHMAN, A., AWAN, M. J., VITKUTĖ-ADŽGAUSKIENĖ, D., DAMAŠEVIČIUS, R., and BAHAJ, S. A. Harris Hawks Sparse Auto-Encoder Networks for Automatic Speech Recognition System. *Applied Sciences*. 2022, vol. 12, no. 3. ISSN 2076-3417. Available from DOI: 10.3390/app12031091.

112. BUTKEVIČIŪTĖ, E., BIKULČIENĖ, L., SIDEKERSKIENĖ, T., BLAŽAUSKAS, T., MASKELIŪNAS, R., DAMAŠEVIČIUS, R., and WEI, W. Removal of Movement Artefact for Mobile EEG Analysis in Sports Exercises. *IEEE Access*. 2019, vol. 7, pp. 7206–7217. Available from DOI: 10.1109/ACCESS.2018.2890335.

113. DAMAŠEVIČIUS, R., NAPOLI, C., SIDEKERSKIENĖ, T., and WOŹNIAK, M. IMF mode demixing in EMD for jitter analysis. *Journal of Computational Science*. 2017, vol. 22, pp. 240–252. ISSN 1877-7503. Available from DOI: 10.1016/j.jocs.2017.04.008.

114. PICAUT, J., CAN, A., FORTIN, N., ARDOUIN, J., and LAGRANGE, M. Low-Cost Sensors for Urban Noise Monitoring Networks—A Literature Review. *Sensors*. 2020, vol. 20, no. 8. ISSN 1424-8220. Available from DOI: 10.3390/s20082256.

115. KANTOVÁ, R. Evaluation of Construction Site Noise to Allow the Optimisation of Construction Processes and Construction Machinery Selection. *Applied Sciences*. 2021, vol. 11, no. 10. ISSN 2076-3417. Available from DOI: 10.3390/app11104389.

116. AHMED, S. S., and GADELMOULA, A. M. Industrial noise monitoring using noise mapping technique: a case study on a concrete block-making factory. *International Journal of Environmental Science and Technology*. 2022, vol. 19, no. 2, pp. 851–862. Available also from: https://doi.org/10.1007/s13762-020-02982-9.

117. LV, Y., LIU, Y., JING, W., WOŹNIAK, M., DAMAŠEVIČIUS, R., SCHERER, R., and WEI, W. Quality Control of the Continuous Hot Pressing Process of Medium Density Fiberboard Using Fuzzy Failure Mode and Effects Analysis. *Applied Sciences*. 2020, vol. 10, no. 13. ISSN 2076-3417. Available from DOI: 10.3390/app10134627.

118. ARAÚJO ALVES, J., NETO PAIVA, F., TORRES SILVA, L., and REMOALDO, P. Low-Frequency Noise and Its Main Effects on Human Health—A Review of the Literature between 2016 and 2019. *Applied Sciences*. 2020, vol. 10, no. 15, p. 5205. Available from DOI: 10.3390/app10155205.

119. PAAR, R., MARENDIĆ, A., JAKOPEC, I., and GRGAC, I. Vibration Monitoring of Civil Engineering Structures Using Contactless Vision-Based Low-Cost IATS Prototype. *Sensors*. 2021, vol. 21, no. 23. ISSN 1424-8220. Available from DOI: 10.3390/s21237952.

120. STANSFELD, S., HAINES, M., and BROWN, B. Noise and Health in the Urban Environment. *Reviews on Environmental Health*. 2000, vol. 15, no. 1-2, pp. 43–82. Available from DOI: doi:10.1515/REVEH.2000.15.1-2.43.

121. MENG, Q., and KANG, J. Effect of sound-related activities on human behaviours and acoustic comfort in urban open spaces. *Science of The Total Environment*. 2016, vol. 573, pp. 481–493. ISSN 0048-9697. Available from DOI: https://doi.org/10.1016/j.scitotenv.2016.08.130.

122. PARK, S. H., LEE, P., and LEE, B. K. Levels and sources of neighbour noise in heavyweight residential buildings in Korea. *Applied Acoustics*. 2017, vol. 120, pp. 148–157. Available from DOI: https://doi.org/10.1016/j.apacoust.2017.01.012.

123. PIJANOWSKI, B. C., VILLANUEVA-RIVERA, L. J., DUMYAHN, S. L., FARINA, A., KRAUSE, B. L., NAPOLETANO, B. M., GAGE, S. H., and PIERETTI, N. Soundscape Ecology: The Science of Sound in the Landscape. *BioScience*. 2011, vol. 61, no. 3, pp. 203–216. ISSN 0006-3568. Available from DOI: 10.1525/bio.2011.61.3.6.

124. QURTHOBI, A., MASKELIŪNAS, R., and DAMAŠEVIČIUS, R. Detection of Mechanical Failures in Industrial Machines Using Overlapping Acoustic Anomalies: A Systematic Literature Review. *Sensors*. 2022, vol. 22, no. 10. ISSN 1424-8220. Available from DOI: 10.3390/s22103888.

125. BHUIYAN, M. R., and UDDIN, J. Deep Transfer Learning Models for Industrial Fault Diagnosis Using Vibration and Acoustic Sensors Data: A Review. *Vibration*. 2023, vol. 6, no. 1, pp. 218–238. ISSN 2571-631X. Available from DOI: 10.3390/vibration6010014.

126. SHARMA, S., SATO, K., and GAUTAM, B. P. A Methodological Literature Review of Acoustic Wildlife Monitoring Using Artificial Intelligence Tools and Techniques. *Sustainability*. 2023, vol. 15, no. 9. ISSN 2071-1050. Available from DOI: https://doi.org/10.3390/su15097128.

127. BROCKMANN-BAUSER, M. How Well Will AI Help Recognize Voice Disorders? A State-of-the-art Review of Current Acoustic Assessment Strategies and Future Applications. *World Journal of Otorhinolaryngology - Head and Neck Surgery*. 2025, vol. n/a, no. n/a. Available from DOI: https://doi.org/10.1002/wjo2.70015.

128. MANIKANDAN, V., and NEETHIRAJAN, S. AI-Driven Bioacoustics in Poultry Farming: A Critical Systematic Review on Vocalization Analysis for Stress and Disease Detection. *Preprints*. 2025. Available from DOI: 10.20944/preprints202505.1369.v1.

129. SEBASTIÁN-GONZÁLEZ, E., and PÉREZ-GRANADOS, C. Geographic Variation in Acoustic Signals in Wildlife: A Systematic Review. *Journal of Biogeography*. 2025, vol. 52, no. 6, e15116. Available from DOI: https://doi.org/10.1111/jbi.15116. e15116 JBI-24-0339.R1.

130. SHOKOUHMAND, S., BHATT, S., and FAEZIPOUR, M. Artificial Intelligence in Respiratory Health: A Review of AI-Driven Analysis of Oral and Nasal Breathing Sounds for Pulmonary Assessment. *Electronics*. 2025, vol. 14, no. 10. ISSN 2079-9292. Available from DOI: 10.3390/electronics14101994.

131. KOIZUMI, Y., SAITO, S., UEMATSU, H., KAWACHI, Y., and HARADA, N. Unsupervised Detection of Anomalous Sound Based on Deep Learning and the Neyman–Pearson Lemma. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2019, vol. 27, no. 1, pp. 212–224. Available from DOI: 10.1109/TASLP.2018.2877258.

132. TANABE, R., PUROHIT, H., DOHI, K., ENDO, T., NIKAIDO, Y., NAKAMURA, T., and KAWAGUCHI, Y. *MIMII DUE: Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection with Domain Shifts due to Changes in Operational and Environmental Conditions*. 2021. Available from arXiv: 2105.02702 [cs.SD].

133. THOIDIS, I., GIOUVANAKIS, M., and PAPANIKOLAOU, G. Semi-Supervised Machine Condition Monitoring by Learning Deep Discriminative Audio Features. *Electronics*. 2021, vol. 10, p. 2471. Available from DOI: 10.3390/electronics10202471.

134. ZAMAN, K., SAH, M., DIREKOGLU, C., and UNOKI, M. A Survey of Audio Classification Using Deep Learning. *IEEE Access*. 2023, vol. 11, pp. 106620–106649. Available from DOI: 10.1109/ACCESS.2023.3318015.

135. SAKTHIVEL, R., PERUMAL KAMALAKANNAN, L., SHANMUGAM, R., and VENUGOPAL, V. Nonauditory Impacts of Industrial Noise Exposures: A Case Study From a Steel Manufacturing Industry. *Safety and Health at Work*. 2025. ISSN 2093-7911. Available from DOI: https://doi.org/10.1016/j.shaw.2025.01.004.

136. TOKER, O. G., ELIBOL, N. T., KURU, E., GORMEZOGLU, Z., GORENER, A., and TOKER, K. Industrial noise: impacts on workers' health and performance below permissible limits. *BMC Public Health*. 2025, vol. 25. Available from DOI: 10.1186/s12889-025-22732-1.

137. LIN, Y.-K., SU, M.-C., and HSIEH, Y.-Z. The Application and Improvement of Deep Neural Networks in Environmental Sound Recognition. *Applied Sciences*. 2020, vol. 10, no. 17. ISSN 2076-3417. Available from DOI: 10.3390/app10175965.

138. ZHOU, X., and ZHAO, M. *Transformer-based Environmental Sound Classification Modeling by Jointing Multi-class Classification and Similarity Clustering*. 2022. Available from DOI: https://doi.org/10.21203/rs.3.rs-2158071/v1.

139. LI, M., HUANG, W., and ZHANG, T. Attention Based Convolutional Neural Network with Multi-frequency Resolution Feature for Environment Sound Classification. *Neural Processing Letters*. 2022, vol. 55. Available from DOI: 10.1007/s11063-022-11041-y.

140. GONG, X., DUAN, H., YANG, Y., TAN, L., WANG, J., and VASILAKOS, A. Improving Audio Classification Method by Combining Self-Supervision with Knowledge Distillation. *Electronics*. 2023, vol. 13, p. 52. Available from DOI: 10.3390/electronics13010052.

141. BOUAZIZ, W., EL-MHAMDI, E.-M., and USUNIER, N. *Targeted Data Poisoning for Black-Box Audio Datasets Ownership Verification*. 2025. Available from arXiv: 2503.10269 [cs.CR].

142. DOSER, J. W., FINLEY, A. O., KASTEN, E. P., and GAGE, S. H. *Assessing soundscape disturbance through hierarchical models and acoustic indices: a case study on a shelterwood logged northern Michigan forest*. 2019. Available from arXiv: 1911.03278 [stat.AP].

143. RANMAL, D., RANASINGHE, P., PARANAYAPA, T., MEEDENIYA, D., and PERERA, C. ESC-NAS: Environment Sound Classification Using Hardware-Aware Neural Architecture Search for the Edge. *Sensors*. 2024, vol. 24, p. 3749. Available from DOI: 10.3390/s24123749.

144. XU, S., and CHEN, Y. Sound classification with time-frequency features in forest environment. *Journal of Physics: Conference Series*. 2024, vol. 2756, p. 012001. Available from DOI: 10.1088/1742-6596/2756/1/012001.

145. PARANAYAPA, T., RANASINGHE, P., RANMAL, D., MEEDENIYA, D., and PERERA, C. A Comparative Study of Preprocessing and Model Compression Techniques in Deep Learning for Forest Sound Classification. *Sensors*. 2024, vol. 24, no. 4. ISSN 1424-8220. Available from DOI: 10.3390/s24041149.

146. AHMAD, M., HASSAN, A., JAVED, S., AHMAD, R., QAZI, S., and ALAM, M. M. Enhanced Forest Sound Classification Dataset: EFSC-24. In: *2024 5th International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering (ELECOM)*. 2024, pp. 1–6. Available from DOI: 10.1109/ELECOM63163.2024.10892163.

147. CHASMAI, M., SHEPARD, A., MAJI, S., and HORN, G. V. *The iNaturalist Sounds Dataset*. 2025. Available from arXiv: 2506.00343 [cs.SD].

148. QURTHOBI, A., DAMAŠEVIČIUS, R., BARZDAITIS, V., and MASKELIUNAS, R. Robust Forest Sound Classification Using Pareto-Mordukhovich Optimized MFCC in Environmental Monitoring. *IEEE Access*. 2025, vol. PP, pp. 1–1. Available from DOI: 10.1109/ACCESS.2025.3535796.

149. ZHUANG, C., LU, Z., WANG, Y., XIAO, J., and WANG, Y. *ACDNet: Adaptively Combined Dilated Convolution for Monocular Panorama Depth Estimation*. 2022. Available from arXiv: 2112.14440 [cs.CV].

150. MARCOT, B., and HANEA, A. What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis? *Computational Statistics*. 2021, vol. 36. Available from DOI: 10.1007/s00180-020-00999-9.

151. NANNI, L., MAGUOLO, G., BRAHNAM, S., and PACI, M. An Ensemble of Convolutional Neural Networks for Audio Classification. *Applied Sciences*. 2021, vol. 11, no. 13, p. 5796. ISSN 2076-3417. Available from DOI: 10.3390/app11135796.

152. ZHANG, Z., XU, S., QIAO, T., ZHANG, S., and CAO, S. *Attention based Convolutional Recurrent Neural Network for Environmental Sound Classification*. 2019. Available from arXiv: 1907.02230 [cs.SD].

153. AYANKOSO, S., WANG, Z., SHI, D., YANG, W., VIKIRU, A., KAMAU, S., MUCHIRI, H., and GU, F. Development of Long-Range, Low-Powered and Smart IoT Device for Detecting Illegal Logging in Forests. *Journal of Dynamics, Monitoring and Diagnostics*. 2024, vol. 3, no. 3, pp. 190–198. Available from DOI: 10.37965/jdmd.2024.550.

154. WOLF-MONHEIM, F. *Spectral and Rhythm Features for Audio Classification with Deep Convolutional Neural Networks*. 2024. Available from arXiv: 2410.06927 [cs.SD].

155. CHOWDHURY, J. H., RAMANNA, S., and KOTECHA, K. Speech emotion recognition with light weight deep neural ensemble model using hand crafted features. *Scientific Reports*. 2025, vol. 15, no. 1, p. 11824. Available from DOI: 10.1038/s41598-025-95734-z.

156. MCFEE, B., RAFFEL, C., LIANG, D., ELLIS, D. P., MCVICAR, M., BATTENBERG, E., and NIETO, O. librosa: Audio and Music Signal Analysis in Python. In: HUFF, K., and BERGSTRA, J. (eds.). *Proceedings of the 14th Python in Science Conference*. 2015, pp. 18–24. Available from DOI: 10.25080/Majora-7b98e3ed-003.

157. MÜLLER, M., KURTH, F., and CLAUSEN, M. Audio Matching via Chroma-Based Statistical Features. In: *ISMIR 2005, 6th International Conference on Music Information Retrieval, London, UK, 11-15 September 2005, Proceedings*. 2005, pp. 288–295. Available also from: http://ismir2005.ismir.net/proceedings/1019.pdf.

158. CHOI, K., FAZEKAS, G., SANDLER, M., and CHO, K. *Convolutional Recurrent Neural Networks for Music Classification*. 2016. Available from arXiv: 1609.04243 [cs.NE].

159. PONS, J., and SERRA, X. Designing efficient architectures for modeling temporal features with convolutional neural networks. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017, pp. 2472–2476. Available from DOI: 10.1109/ICASSP.2017.7952601.

160. ZHAO, X., and WANG, D. Analyzing noise robustness of MFCC and GFCC features in speaker identification. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013, pp. 7204–7208. Available from DOI: 10.1109/ICASSP.2013.6639061.

161. CHO, T., and BELLO, J. P. On the relative importance of individual components of chord recognition systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2014, vol. 22, no. 3, pp. 477–492. Available from DOI: 10.1109/TASLP.2013.2295926.

162. BITTNER, R. M., SALAMON, J., TIERNEY, M., MAUCH, M., CANNAM, C., and BELLO, J. P. MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research. In: *15th International Society for Music Information Retrieval Conference*. Taipei, Taiwan, 2014, pp. 155–160. Available from DOI: 10.5281/zenodo.1417889.

163. HARTE, C., SANDLER, M., and GASSER, M. Detecting harmonic change in musical audio. In: *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*. ACM, 2006, pp. 21–26. Available from DOI: 10.1145/1178723.1178727.

164. MCVICAR, M., FREEMAN, T., and DE BIE, T. Mining the correlation between lyrical and audio features and the emergence of mood. In: *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*. 2011, pp. 783–788.

165. ISLAM, M., and ALI, M. N. Y. Environmental Sound Classification Using Feature Fusion of MFCCs, Mel-spectrogram, and Chroma. In: *2024 27th International Conference on Computer and Information Technology (ICCIT)*. 2024, pp. 3212–3217. Available from DOI: 10.1109/ICCIT64611.2024.11021738.

166. CAO, K., ZHANG, T., and HUANG, J. Advanced hybrid LSTM-transformer architecture for real-time multi-task prediction in engineering systems. *Scientific Reports*. 2024, vol. 14. Available from DOI: 10.1038/s41598-024-55483-x.

167. YADAV, H., SHAH, P., GANDHI, N., VYAS, T., NAIR, A., DESAI, S., GOHIL, L., TANWAR, S., SHARMA, R., MARINA, V., and RABOACA, M. S. CNN and Bidirectional GRU-Based Heartbeat Sound Classification Architecture for Elderly People. *Mathematics*. 2023, vol. 11, no. 6. ISSN 2227-7390. Available from DOI: 10.3390/math11061365.

168. ETIENNE, C., FIDANZA, G., PETROVSKII, A., DEVILLERS, L., and SCHMAUCH, B. CNN+LSTM Architecture for Speech Emotion Recognition with Data Augmentation. In: *Workshop on Speech, Music and Mind (SMM 2018)*. ISCA, 2018. Available from DOI: 10.21437/smm.2018-5.

169. ZHANG, Y., and MARTÍNEZ-GARCÍA, M. Machine Hearing for Industrial Fault Diagnosis. In: *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*. 2020, pp. 849–854. Available from DOI: 10.1109/CASE48305.2020.9216787.

170. INAPAGOLLA, R. K., and BABU, K. . K. Audio Fingerprinting to Achieve Greater Accuracy and Maximum Speed with Multi-Model CNN-RNN-LSTM in Speaker Identification: Speed with Multi-Model CNN-RNN-LSTM in Speaker Identification. *International Journal of Computational and Experimental Science and Engineering*. 2025, vol. 11. Available from DOI: 10.22399/ijcesen.1138.

171. KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*. Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc., 1995, pp. 1137–1143. IJCAI'95. ISBN 1558603638.

172. ZHOU, Y., LONG, Y., and WEI, H. Acoustic-Sensing-Based Attribute-Driven Imbalanced Compensation for Anomalous Sound Detection without Machine Identity. *Sensors*. 2023, vol. 23, no. 21. ISSN 1424-8220. Available from DOI: 10.3390/s23218984.

173. ZHANG, Z., XU, S., CAO, S., and ZHANG, S. *Deep Convolutional Neural Network with Mixup for Environmental Sound Classification*. 2018. Available from arXiv: 1808.08405 [cs.SD].

174. DEMIR, F., ABDULLAH, D. A., and SENGUR, A. A New Deep CNN Model for Environmental Sound Classification. *IEEE Access*. 2020, vol. 8, pp. 66529–66537. Available from DOI: 10.1109/ACCESS.2020.2984903.

175. CORNEANU, C., MADADI, M., ESCALERA, S., and MARTINEZ, A. *Computing the Testing Error without a Testing Set*. 2020. Available from arXiv: 2005.00450 [cs.CV].

176. HASTIE, T., TIBSHIRANI, R., and FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer, 2009. ISBN 978-0-387-84857-0. Available from DOI: 10.1007/978-0-387-84858-7.

177. POWERS, D. M. W. *Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation*. 2020. Available from arXiv: 2010.16061 [cs.LG].

178. SOKOLOVA, M., and LAPALME, G. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*. 2009, vol. 45, no. 4, pp. 427–437. ISSN 0306-4573. Available from DOI: https://doi.org/10.1016/j.ipm.2009.03.002.

179. FAWCETT, T. An introduction to ROC analysis. *Pattern Recognition Letters*. 2006, vol. 27, no. 8, pp. 861–874. ISSN 0167-8655. Available from DOI: https://doi.org/10.1016/j.patrec.2005.10.010. ROC Analysis in Pattern Recognition.

180. PROVOST, F., and FAWCETT, T. *Robust Classification for Imprecise Environments*. 2000. Available from arXiv: cs/0009007 [cs.LG].

181. RICHARDSON, E., TREVIZANI, R., GREENBAUM, J. A., CARTER, H., NIELSEN, M., and PETERS, B. The receiver operating characteristic curve accurately assesses imbalanced datasets. *Patterns*. 2024, vol. 5, no. 6, p. 100994. ISSN 2666-3899. Available from DOI: https://doi.org/10.1016/j.patter.2024.100994.

182. MAATEN, L. van der, and HINTON, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research*. 2008, vol. 9, no. 86, pp. 2579–2605. Available also from: http://jmlr.org/papers/v9/vandermaaten08a.html.

183. ANDAYANI, F., THENG, L. B., TSUN, M. T., and CHUA, C. Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files. *IEEE Access*. 2022, vol. 10, pp. 36018–36027. Available from DOI: 10.1109/ACCESS.2022.3163856.

184. LORENA, A. C., GARCIA, L. P. F., LEHMANN, J., SOUTO, M. C. P., and HO, T. K. *How Complex is your classification problem? A survey on measuring classification complexity*. 2020. Available from arXiv: 1808.03591 [cs.LG].

185. EWERT, S., MULLER, M., and GROSCHE, P. High resolution audio synchronization using chroma onset features. In: *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2009, pp. 1869–1872. Available from DOI: 10.1109/ICASSP.2009.4959972.

186. MAUCH, M., and DIXON, S. Simultaneous Estimation of Chords and Musical Context From Audio. *IEEE Transactions on Audio, Speech, and Language Processing*. 2010, vol. 18, no. 6, pp. 1280–1289. Available from DOI: 10.1109/TASL.2009.2032947.

187. PEETERS, G., and RICHARD, G. Deep Learning for Audio and Music. In: BENOIS-PINEAU, J., and ZEMMARI, A. (eds.). *Multi-faceted Deep Learning: Models and Data*. Cham: Springer International Publishing, 2021, pp. 231–266. ISBN 978-3-030-74478-6. Available from DOI: 10.1007/978-3-030-74478-6_10.

188. SLAM, W., LI, Y., and UROUVAS, N. Frontier Research on Low-Resource Speech Recognition Technology. *Sensors*. 2023, vol. 23, no. 22. ISSN 1424-8220. Available also from: https://www.mdpi.com/1424-8220/23/22/9096.

189. CHEN, K., DU, X., ZHU, B., MA, Z., BERG-KIRKPATRICK, T., and DUBNOV, S. *HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection*. 2022. Available from arXiv: 2202.00874 [cs.SD].

190. CHEN, S., WU, Y., WANG, C., LIU, S., TOMPKINS, D., CHEN, Z., and WEI, F. *BEATs: Audio Pre-Training with Acoustic Tokenizers*. 2022. Available from arXiv: 2212.09058 [eess.AS].

191. ZAMAN, K., LI, K., SAH, M., DIREKOGLU, C., OKADA, S., and UNOKI, M. Transformers and audio detection tasks: An overview. *Digital Signal Processing*. 2025, vol. 158, p. 104956. ISSN 1051-2004. Available from DOI: https://doi.org/10.1016/j.dsp.2024.104956.

192. DING, S., ZHANG, S., and YANG, C. Machine tool fault classification diagnosis based on audio parameters. *Results in Engineering*. 2023, vol. 19, p. 101308. ISSN 2590-1230. Available from DOI: 10.1016/j.rineng.2023.101308.

193. SIRAJ, F. M., AYON, S. T. K., and UDDIN, J. A Few-Shot Learning Based Fault Diagnosis Model Using Sensors Data from Industrial Machineries. *Vibration*. 2023, vol. 6, no. 4, pp. 1004–1029. ISSN 2571-631X. Available from DOI: 10.3390/vibration6040059.

194. PU, H., WEN, Z., SUN, X., HAN, L., NA, Y., LIU, H., and LI, W. Research on the mechanical fault diagnosis method based on sound signal and IEMD-DDCNN. *International Journal of Intelligent Computing and Cybernetics*. 2023, vol. 16. Available from DOI: 10.1108/IJICC-09-2022-0253.

195. ALAGELE, Z., ALKAFAJE, S., and JABAR, R. Designing a Deep Autoencoder Neural Network for Detecting Sound Anomalies in Smart Factories Using Unsupervised Learning. *BIO Web of Conferences*. 2024, vol. 97, p. 00027. Available from DOI: 10.1051/bioconf/20249700027.

196. CHANDRAKALA, S., PIDIKITI, A., and MAHATHI, P. Spectro Temporal Fusion with CLSTM-Autoencoder based approach for Anomalous Sound Detection. *Neural Processing Letters*. 2024, vol. 56. Available from DOI: 10.1007/s11063-024-11485-4.

197. SARKAR, P., and ETEMAD, A. *Self-Supervised Audio-Visual Representation Learning with Relaxed Cross-Modal Synchronicity*. 2022. Available from arXiv: 2111.05329 [cs.CV].

198. CHEN, X., WANG, M., KAN, R., and QIU, H. Improved Patch-Mix Transformer and Contrastive Learning Method for Sound Classification in Noisy Environments. *Applied Sciences*. 2024, vol. 14, p. 9711. Available from DOI: 10.3390/app14219711.

199. SIMIYU, D., MUCHIRI, H., VIKIRU, A., BUTIME, J., and MUTI, Z. A Forest Acoustics – Temporal Frequency Convolution Neural Network Model for Detecting Illegal Logging Activities in Forest. In: *2024 5th International Conference on Smart Sensors and Application (ICSSA)*. 2024, pp. 1–6. Available from DOI: 10.1109/ICSSA62312.2024.10788639.

200. SIMS, Y., MENDES, A., and CHALUP, S. *Embedding-Space Diffusion for Zero-Shot Environmental Sound Classification*. 2025. Available from arXiv: 2412 . 03771 [cs.SD].

201. JAHANGIR, R., NAUMAN, M. A., ALROOBAEA, R., ALMOTIRI, J., MALIK, M. M., and ALZAHRANI, S. M. Deep Learning-based Environmental Sound Classification Using Feature Fusion and Data Enhancement. *Computers, Materials & Continua*. 2023, vol. 74, no. 1, pp. 1069–1091. ISSN 1546-2226. Available from DOI: 10.32604/cmc.2023.032719.

CURRICULUM VITAE AND DESCRIPTION OF CREATIVE ACTIVITIES (CV)

Personal Details

Name, Surname	:	Ahmad, Qurthobi
Date of Birth	:	February 2, 1985
E-mail	:	ahmad.qurthobi@ktu.lt

Educations

2003 – 2007	:	**Bachelor of Engineering in Telecommunication** Sekolah Tinggi Teknologi Telkom (now Telkom University), Bandung, Indonesia
2008 – 2011	:	**Master of Science in Instrumentation and Control** Bandung Institute of Technology, Bandung, Indonesia
2021 – Present	:	**Doctoral Candidate of Philosophy in Informatics Engineering** Kaunas University of Technology

Professional Experiences

2012 – 2012	:	**PT D&C Engineering Company** Junior Instrumentation and control engineer
2013 – Present	:	**Telkom University** Lecturer in the Department of Engineering Physics, School of Electrical Engineering

Area of Research Interests

- Artificial Intelligence
- Control System Engineering
- Deep Learning
- Environmental Acoustics
- Electrical Machines
- Instrumentation System
- Machine Learning
- Power Electronics
- Renewable Energy
- System Dynamics

LIST OF SCIENTIFIC PAPERS AND SCIENTIFIC CONFERENCES

Articles indexed in Web of Science and Scopus

1. Qurthobi, Ahmad; Maskeliūnas, Rytis; Damaševičius, Robertas. Detection of mechanical failures in industrial machines using overlapping acoustic anomalies: a systematic literature review // Sensors. Basel : MDPI. ISSN 1424-8220. 2022, vol. 22, iss. 10, art. no. 3888, p. 1-20. DOI: 10.3390/s22103888. [Science Citation Index Expanded (Web of Science); Scopus; MEDLINE] [IF: 3.900; AIF: 4.333; IF/AIF: 0.900; Q2 (2022, InCites JCR SCIE)] [Field: T 007] [Contribution: 0.334]
2. Qurthobi, Ahmad; Damaševičius, Robertas; Barzdaitis, Vytautas; Maskeliūnas, Rytis. Robust forest sound classification using Pareto-Mordukhovich optimized MFCC in environmental monitoring // IEEE Access. Piscataway, NJ : IEEE. ISSN 2169-3536. 2025, vol. 13, p. 20923-20944. DOI: 10.1109/ACCESS.2025.3535796. [Science Citation Index Expanded (Web of Science); Scopus] [IF: 3.600; AIF: 4.533; IF/AIF: 0.794; Q2 (2024, InCites JCR SCIE)] [Field: T 007, N 009] [Contribution: 0.250]

International conference proceedings

1. The effect of augmentation and filtration on noisy environment's acoustic signals to detect abnormalities in industrial machines based on artificial neural networks / Ahmad Qurthobi, Rytis Maskeliūnas.
   Qurthobi, Ahmad, author.; Maskeliūnas, Rytis, author.; "Elsevier" group publisher.
   DOI 10.1016/j.procs.2023.03.068; SCOPUS EID 2-s2.0-85164525267;
   SCOPUS PII S1877050923006038; SCOPUS ID 85164525267; OA 1; SOURCE OA 1; Dimensions ID 1157303934; ISSN/eISSN 1877-0509; 2023
   Procedia computer science: 14th international conference on ambient systems, networks and technologies networks, ANT 2023 and the 6th international conference on emerging data and industry 4.0, EDI40 2023 / edited by Elhadi Shakshuki. Amsterdam: Elsevier 2023, vol. 220
2. Deep learning and acoustic approach for mechanical failure detection in industrial machinery / Ahmad Qurthobi, Rytis Maskeliūnas.
   Qurthobi, Ahmad, author.; Maskeliūnas, Rytis, author.; IOP (Institute of Physics) Publisher.
   DOI 10.1088/1742-6596/2673/1/012032; SCOPUS EID 2-s2.0-85182267301; SCOPUS ID 85182267301; OA 1; Dimensions ID 1167370942; ISSN/eISSN 1742-6588; eISSN 1742-6596; 2023
   Journal of physics: conference series: 4th engineering physics international conference 2023 (EPIC 2023), 26-27 September 2023, Bandung, Indonesia. IOP publishing 2023, vol. 2673, iss. 1, art. no. 012032 1742-6596
3. A hybrid machine learning model for forest wildfire detection using sounds / Robertas Damasevicius, Ahmad Qurthobi, Rytis Maskeliūnas.
   Damasevicius, Robertas, author.; Qurthobi, Ahmad, author.; Maskeliūnas, Rytis, autorius.; IEEE (Institute of Electrical and Electronics Engineers) publisher.

DOI 10.15439/2024F7263; CROSSREF CITATION ID 166381865; ISBN 9788396960177; DOI 10.15439/978-83-969601-6-0; SCOPUS EID 2-s2.0-85212270457; WOS ID 001413201600011; ISSN/eISSN 2300-5963; ISBN/eISBN 9788396960184; eISBN 9788396960160; 2024
Proceedings of the 19th conference on computer science and intelligence systems (FedCSIS), September 8–11, 2024. Belgrade, Serbia. Piscataway, NJ: IEEE, 2024 9788396960160

ACKNOWLEDGEMENT

Above all, I offer my deepest gratitude to Almighty Allah, the Most Merciful and Compassionate, for granting me the strength, resilience, and patience to overcome difficulties and successfully complete this important chapter of my life. Through His guidance, I remained steadfast, grew through adversity, and reached this meaningful milestone.

I would like to express my sincere appreciation to my supervisor, Prof. Dr. Rytis Maskeliūnas, for his unwavering confidence in me and for the invaluable opportunity to pursue my doctoral studies under his supervision. His guidance, support, and encouragement have been instrumental in shaping my academic journey and have contributed significantly to both my professional and personal growth.

I am also deeply grateful to Prof. Dr. Robertas Damaševičius for the opportunity to engage in meaningful research at the Center of Real-Time Computer Systems and within the Forest 4.0 initiative. His expertise and constructive feedback have played a key role in strengthening my research skills and shaping this dissertation.

Furthermore, I extend my sincere thanks to the esteemed reviewers (Prof. Dr. Egidijus Kazanavičius, Prof. Dr. Tomas Blažauskas, Prof. Dr. Renaldas Urniežius, Prof. Dr. Nikolaj Goronin, Prof. Dr. Renaldas Raišutis, Prof. Dr. Dimitrij Šešok, and Assoc. Prof. Zenun Kastrati) whose insightful comments and thorough evaluations have significantly improved the quality, rigor, and clarity of this PhD thesis.

I would like to express my sincere gratitude to Kaunas University of Technology for fostering an outstanding academic environment, providing committed and insightful supervision, and ensuring access to the key facilities and resources that were essential for the successful completion of this work. I am equally thankful to the Research Council of Lithuania (LMT) for its crucial financial support, which not only enabled the continuity of my research activities but also significantly contributed to the growth of my academic competencies and overall professional development.

I am deeply grateful to Forest 4.0 for the opportunity to expand my expertise in artificial intelligence, particularly in forest monitoring. This experience has strengthened both my theoretical understanding and practical approach to real-world environmental challenges. The research presented in this dissertation was supported by the Horizon Europe Framework Programme (HORIZON) under the Teaming for Excellence call (HORIZON-WIDERA-2022-ACCESS-01-two-stage), through the Centre of Excellence in Smart Forestry “Forest 4.0” (No. 101059985). I sincerely appreciate the collaborative environment, intellectual exchange, and support provided by the consortium and the European Union.

I also extend my gratitude to Telkom University for enabling me to pursue my studies at Kaunas University of Technology. Their continuous moral and financial support has been essential to my academic and professional development, and I look forward to contributing to the Physics Engineering Study Program and the broader academic community upon my return.

My deepest appreciation goes to my family for their unwavering support. I am especially grateful to my mother, Siti Hamdah, whose encouragement, prayers, and guidance have been invaluable throughout my journey.

I owe deep gratitude to my dear wife, Sayyidah Adiyah Lailah Mubarakah Azhimah (SALMA), whose unwavering love, patience, and encouragement have been a constant source of strength. Her quiet endurance, her many sacrifices, and her steadfast belief in me have carried me through every challenge of this journey and made it possible for me to persevere and grow.

To my dearest son, Abdullah Ahmad Ghazalie, my heart overflows with gratitude for the happiness and light you bring into my life each day. I also want to express my heartfelt apology for the times I was unable to be with you during your earliest years. My absence was never a reflection of my love for you, and I hope that, as you grow older, you will come to understand the reasons and know that you have always been, and will always be, deeply cherished.

Finally, I extend my heartfelt appreciation to my friends in Kaunas, in particular the KTU students and alumni (Dr. Sarmad Maqsood, Dr. Jewel Sengupta, Musyyab Yousufi, Gopi Kompelli, and many others) as well as the Indonesian community in Kaunas (Rhevin F. Putra, Muammar I. Aulia, Rendy Anggriawan, Alvin D. Nugroho, and others). Their support and companionship have made this journey both meaningful and memorable.

UDK 004.032.26+681.88](043.3)

SL344. 2026-03-18, 23,25 leidyb. apsk. l. Tiražas 14 egz. Užsakymas 035.
Išleido Kauno technologijos universitetas, K. Donelaičio g. 73, 44249 Kaunas
Spausdino leidyklos „Technologija“ spaustuvė, Studentų g. 54, 51424 Kaunas