An Optimised Grid Search Based Framework for Robust Large-Scale Natural Soundscape Classification

Thomas Napier(✉), Euijoon Ahn, Slade Allen-Ankins, and Ickjai Lee

James Cook University, Townsville, QLD 4811, Australia
{thomas.napier,euijoon.ahn,slade.allenankins,ickjai.lee}@jcu.edu.au

Abstract. Large-scale natural soundscapes are remarkably complex and offer invaluable insights into the biodiversity and health of ecosystems. Recent advances have shown promising results in automatically classifying the sounds captured using passive acoustic monitoring. However, the accuracy performance and lack of transferability across diverse environments remains a challenge. To rectify this, we propose a robust and flexible ecoacoustics sound classification grid search-based framework using optimised machine learning algorithms for the analysis of large-scale natural soundscapes. It consists of four steps: pre-processing including the application of spectral subtraction denoising to two distinct datasets extracted from the Australian Acoustic Observatory, feature extraction using Mel Frequency Cepstral Coefficients, feature reduction, and classification using a grid search approach for hyperparameter tuning across classifiers including Support Vector Machine, k-Nearest Neighbour, and Artificial Neural Networks. With 10-fold cross validation, our experimental results revealed that the best models obtained a classification accuracy of 96% and above in both datasets across the four major categories of sound (biophony, geophony, anthrophony, and silence). Furthermore, cross-dataset validation experiments using a pooled dataset highlight that our framework is rigorous and adaptable, despite the high variance in possible sounds at each site.

Keywords: Ecoacoustics · Signal Processing · Machine Learning · Optimised Grid Search

1 Introduction

Australia is one of the most biodiverse regions on Earth, yet many species are under threat [2]. Effective monitoring solutions and techniques have now become imperative for the tracking of at-risk species. Ecoacoustics serves as one such solution, which has gained recent attention for its potential in ecological conservation [1,8,16,17]. Leveraging modern advancements in low-cost sound recording and data storage solutions, remote sensor monitoring of natural soundscapes are now possible by way of large-scale Passive Acoustic Monitoring (PAM) [5]. This, in turn, allows ecoacoustics studies to now utilise the wide spatial and temporal soundscape coverage enabled by PAM [14].

T. Liu et al. (Eds.): AI 2023, LNAI 14471, pp. 468–479, 2024.
https://doi.org/10.1007/978-981-99-8388-9_38

In recognition of these conservation benefits, a new large-scale sensor network was established called the Australian Acoustic Observatory (A2O) [14]. The A2O seeks to capture sounds at an ecosystem level. To do so, over 360 listening stations are situated at 90 different sites across Australia to capture sounds from as many ecoregions as possible. Natural soundscapes are broadly composed of four sound groups including: biophony (the sounds produced by animals), geophony (natural non-biological sounds like wind or water), anthrophony (sounds induced by humans) and periods of silence. By segregating these recorded sounds into the four primary categories, researchers can gain a nuanced understanding of their relative balance which is useful for analysing the state of biodiversity and effects of human impact [6,7]. Despite this, much of the existing Machine Learning (ML) and Deep Learning (DL) ecoacoustics research relies on single-species, often non-ecological datasets [4,5,8,15]. While these are valuable, they don't provide the comprehensive insights needed for broader ecological conclusions. Analysing large-scale PAM datasets like those derived from the A2O, are uniquely challenging for several reasons. Firstly, they feature an exceptionally high variance in the types of sounds captured. Natural environments are profoundly complex and dynamic places, which change on a day-to-day basis. This is compounded by the sheer diversity of species calls, which often overlap each other in both time and frequency [13]. To address these intricacies, this research will explore the integration of select Feature Extraction (FE), Feature Reduction (FR), and denoising techniques. Our framework will need to be flexible and robust with these considerations in mind, while still maintaining high accuracy. We believe that using a grid search based approach for both the hyperparameter-sensitive FR techniques, and associated classifiers, will achieve these desired outcomes.

The contributions of this work can be summarised as follows:

- Proposal of a novel grid search based methodology that leverages ML algorithms to identify and categorise distinct sounds within diverse ecological environments;
- Utilisation of a human-inspired approach to feature representation, aimed at improving classification performance;
- Provision of a range of exploratory experiments using the real ecoacoustics data captured from two distinct ecosystems in Australia to evaluate the effectiveness of each classifier-based prediction model on unseen contextual test cases;
- Identification and suggestion of the most suitable algorithms and supervised learning techniques for classifying ecoacoustics sounds into the four broad categories in soundscape ecology.

2 Related Works

Natural soundscapes are inherently complex due to the vast diversity of species calls which can change on a day-to-day basis. For this reason, models must be generalisable with these complications considered in order to be genuinely useful in the real world. However, this is often not the case. Many existing approaches are trained and tested on single species datasets, or ones with low taxonomic variation, which does not accurately reflect the real-world [5]. Those which do use natural soundscape datasets such as one study in 2021 [16], have employed the use of summary statistics in the form of acoustic indices. By using overly summative features like acoustic indices, crucial details are lost which diminishes the richness of the recordings, resulting in a loss of overall accuracy down to 70% in some downstream classification tasks.

Another study using a natural soundscape dataset has shown some merit in classifying broad sound groupings. Here, the authors were able to achieve relatively accurate results across each of the major sound groups (biophony, geophony, and anthrophony), ranging from 88% to 95% accuracy using an Artificial Neural Network (ANN) classifier with Mel Frequency Cepstral Coefficients (MFCCs) [6]. However, the study was conducted using a dataset constructed from four recorders in a comparatively smaller study site compared to the A2O, situated on the border of France and Switzerland. Furthermore, sounds were collected episodically for 1 min, every 15 min, rather than continuously as with recordings derived from the large-scale initiatives like the A2O [14]. Thus, the flexibility of their approach is unknown as the authors did not experiment with data from alternative study sites.

Furthermore, several studies have used overly sanitised or non-representative datasets. For this reason, many approaches are capable of performing well on datasets containing specific species like frogs [3] and birds [8,15], however, they lack flexibility to different ecoregions, and are too specialised to assist in answering broader ecological questions. Accurate, adaptable ecoacoustics classification is pivotal for ecologists. Not only can it facilitate comprehensive biodiversity assessments, but it also plays a crucial role in early detection and prevention of species loss.

3 The Proposed Framework

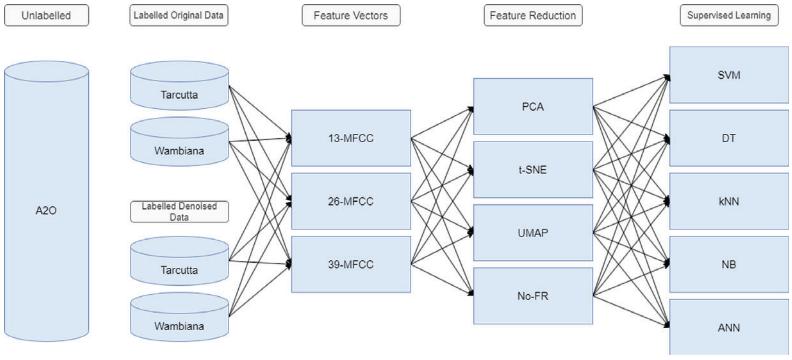
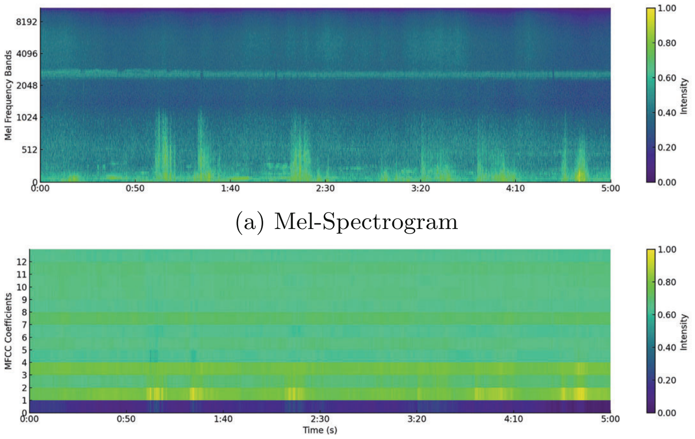


Fig. 1. A conceptual diagram for a robust grid search based framework for use with large-scale ecoacoustics data.

In this study, we propose a novel grid search based classification framework for use with large-scale natural soundscape audio signals based on MFCC feature vectors and ML techniques as shown in Fig. 1. Raw, unlabelled audio was collected directly from the A2O from two geographically different ecoregions to test for cross-site validity. The raw signals were segmented into 4.5-second-long non-overlapping sequences, resulting in a combined total of 8,841 samples, which were subsequently annotated by human experts. From this point we established two datasets: one was unaltered containing the original 8,841 labelled samples, and the second was a direct copy with spectral subtraction denoising applied to all samples. From this, we extracted the MFCC feature vectors as represented by the corresponding heatmap visualised in Fig. 2(b). To achieve this, we compute a Mel-Spectrogram with 128 Mel bands, as seen in Fig. 2(a). This was chosen due to its ability to closely mirror the human auditory system's frequency perception. Furthermore, transforming the spectrograms onto a decibel logarithmic scale has been shown to successfully capture the underlying signal properties. This has been validated across a range of ML tasks including bird song classification [11,19], as well as its usage in acoustic scene classification [6,12,20]. The extraction of 13 MFCC features provides a compact representation of the audio's spectral shape, allowing for high intra-class variability for class discrimination. In addition, we also include the first- and second-order derivatives of MFCCs in a 26- and 39-feature vector, respectively, as a way of capturing the audio's temporal dynamics. Finally, we apply a min-max normalisation scheme to avoid any single feature from disproportionately influencing the model due to its scale. Due to the high-dimensional nature of these feature vectors, we also used several FR techniques including Principle Component Analysis (PCA),



(a) Mel-Spectrogram

(b) MFCC Feature Vector Heatmap

Fig. 2. An example Mel-Spectrogram and associated 13-MFCC feature vector heatmap derived from the same 5 min of biophony audio from the Tarcutta A2O site.

t-distributed Stochastic Neighbour Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) to measure their effect on downstream model performance and to counter potential cases of overfitting. For classifiers, we implemented a grid search approach for tuning the hyperparameters of several ML models including Artificial Neural Network (ANN) and k-Nearest Neighbour (k-NN), and used the default parameter settings for Support Vector Machine (SVM), Decision Tree (DT), and Naïve Bayes (NB). The performance of each classifier was evaluated using macro-based metrics for accuracy, precision, recall, and F1-score, specifically to account for the imbalanced nature of the dataset.

4 Datasets and Experiments

4.1 Datasets

Each A2O site consists of a group of four sensors, two in dry areas and two in wet areas, each recording continuously for 24 h per day [14]. In this study, we selected a two-week period during Australia's Autumn season for further analysis. For each day and sensor, a Long-Duration False Colour (LDFC) spectrogram was generated using a select combination of acoustic indices [18]. LDFC spectrograms provide a snapshot into the day's acoustic activity which can be visually scanned.

Table 1. Tarcutta and Wambiana dataset breakdown by class.

Sound Category	Tarcutta (# samples)	Wambiana (# samples)	Sounds Included
Biophony	2,379	685	Any sound generated by animals (birds, frogs, insects, etc.)
Geophony	2,014	573	Any sound from the earth (water, wind, fire, etc.)
Anthrophony	532	252	Any human-made sound (cars, airplanes, human speech, etc.)
Other/Silence	1,679	727	Mostly represents long periods of silence but can include sounds like ”white noise” or ”pink noise”, electromagnetic interference, etc.
Total	6,604	2,237	8,841

Two datasets were collected by visually analysing the corresponding LDFC spectrogram and aurally listening for each of the main categories of sound. The first was collected from the Tarcutta site, a temperate woodland area located in south-western New South Wales. The second was from Wambiana, a small station in the tropical Far North Queensland located three hours outside of Townsville. We chose these sites specifically because they are geographically spread. This spread ensures that there is a high intra-class variance in the sounds captured, which is representative of the real-world. Furthermore, sounds vary greatly on a day-to-day basis, which we wanted to capture by taking samples from as many days as possible. As seen in Table 1, the final datasets were roughly equally distributed across three of the four major sound groupings (biophony, geophony and other/silence), with less anthrophony due to the seclusion of the sites. Importantly, no preprocessing was conducted prior to manual annotation to allow for the subsequent models to learn from real-world examples, with the noise and variability of environmental factors included.

4.2 Feature Reduction Experiments

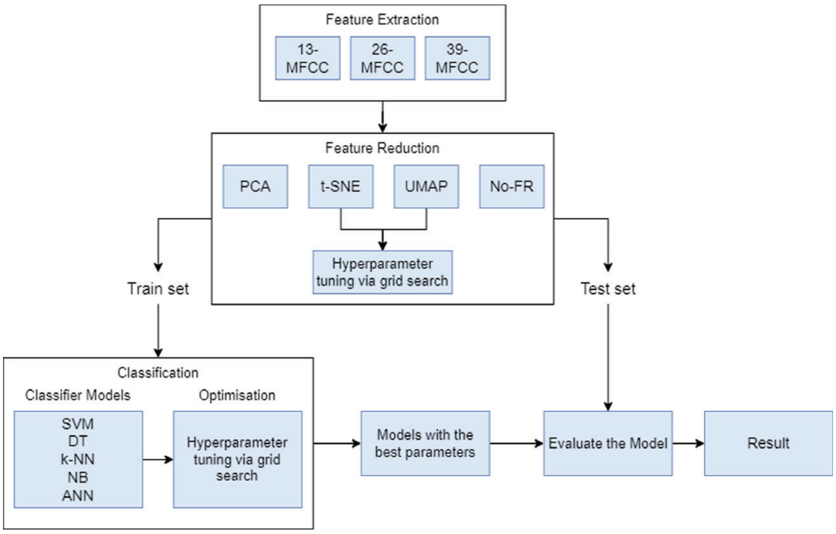


Fig. 3. The implementation workflow of the proposed framework.

Here we present the implementation workflow of the proposed framework as per Fig. 3. Firstly, to improve computational efficiency and mitigate the curse of dimensionality, we employed FR. Three techniques were selected for further analysis including PCA, t-SNE and UMAP, with No-FR as a baseline. A vital part of using PCA in practice is the ability to estimate how many components are needed to describe the data. Similarly, the results of t-SNE and UMAP can vary depending on the choice of hyperparameters, namely perplexity for t-SNE, and number of neighbours and minimum distance for UMAP. For PCA, we employed the cumulative explained variance ratio as a determinant for selecting the optimal number of components. By plotting this ratio against the number of components, we discerned that 7 components were sufficient to capture approximately 95% of the total variance in the data, across all 13, 26 and 39 feature vectors, striking a balance between data compression and information retention.

To find the optimal choice of hyperparameters for t-SNE and UMAP, we applied a grid search approach across both the Tarcutta and Wambiana datasets.

For t-SNE, optimal perplexity values typically fall between 5 and 50 [9]. Perplexity is a parameter designed to shift the attention of local versus global structures in the datasets. A low perplexity emphasises local structure, while a high value may reveal more global structure. For our datasets, we found that a perplexity value of 50 captured enough local distinctions while revealing the broader patterns. Similarly, UMAP relies on the tuning of the number of neighbours and minimum distance parameters to balance local versus global structures in the data embedding [10]. Here, a lower number of neighbours will focus on local structure, and the minimum distance controls how tightly UMAP packs points together, with lower values leading to more compact clusters. We found that a value of 6 and 0.1 for the number of neighbours and minimum distance, respectively, produced a result where clusters representing different sound patterns remained distinct and without excessive overlap or bridging.

4.3 Denoising with Spectral Subtraction

For the next step in our framework, we retained one dataset in its original state to serve as a reference. On a copied version, we applied spectral subtraction. Spectral subtraction is a common approach to audio noise reduction. It functions by generating a noise profile and subtracting it from the original signal. This process preserves vital information like bird and frog calls, while eliminating stationary background noise often found in environmental recordings such as rain [21]. By performing this, we were able to conduct a comparative analysis between the raw and denoised data, thereby understanding its impact on subsequent classification tasks.

4.4 Optimising Classifier Models via Grid Search

To effectively evaluate the classification performance on the uniquely challenging natural soundscape datasets, we employed a diverse set of supervised learning techniques including SVM, DT, k-NN, NB and ANN. To find this selection, we examined several techniques and chose based on their differences in learning principles and foundational algorithms. We wanted to showcase a range of strategies to investigate how they perform with the inherent complexities of the ecoacoustics datasets. SVM is well-known for handling high-dimensional data, making it an ideal candidate for this study. Given its ability to handle high-dimensional data and its efficacy in finding optimal hyperplanes for classification, it was a clear choice. For our purposes, we used the default parameters, as they offer a solid benchmark and are often optimised for a broad range of datasets.

We selected DT because they are interpretable, and their hierarchical structure allows for an intuitive understanding of decision processes. Using default parameters provides a baseline and avoids overfitting that might arise from excessively deep or complex trees. Similarly, NB was incorporated with default parameters to test how the model's underlying probabilistic assumptions perform with these datasets. With a range of algorithmic approaches selected, we identified the need for hyperparameter optimisation, where the model's sensitivity to them significantly influences the outcome. As such, we employed a grid search approach for both k-NN and ANN, as both require a thorough evaluation for performance optimisation. k-NN is particularly effective in situations where data might form natural clusters based on similarity. However, the choice of k is crucial. As such, we utilised a grid search approach to ascertain the optimal number of neighbours k to consider, ranging from (k = 1, 3, 5, 7, ... 31), ensuring our model was neither too generalised nor too specific. Similarly, for ANNs with their inherent flexibility, certain hyperparameters such as solver type and hidden layer sizes required tuning.

5 Results and Discussion

5.1 Classification Results

Experiment Setup. We applied the grid search method with 10-fold cross validation using several datasets: firstly, the unaltered datasets from the Tarcutta and Wambiana sites derived from the A2O containing 6,604 and 2,237 signal samples, respectively, as well as their denoised versions where spectral subtraction was applied. Additionally, we further constructed a combined dataset, pooling samples from both sites together. 10-fold cross validation ensured that each approach minimised the risk of overfitting and provided us with a more reliable assessment of their performance. Furthermore, to ascertain the model's capability for generalisation across different ecoregions, we designed two cross-dataset validation tests. Instead of using 10-fold cross validation, we used an 80%/20% train/test split, respectively. By training on one site, and testing it on another, we could evaluate how well the models adapted to new, unseen data, which is vital in large-scale ecoacoustics, where conditions can greatly vary between sites. We assess model performance using the following key evaluation metrics: accuracy (proportion of correct predictions to total predictions), precision (proportion of true positive predictions to total positive predictions), recall (proportion of true positive predictions to actual positives), and the F1-score (harmonic mean of precision and recall). After conducting several simulations, the classification accuracy performance of each combination of feature vector, tuned feature reduction approach, and optimised classifier for each dataset is showcased in Fig. 4.

Findings. For ANNs, we found the Adam optimiser with a hidden layer configuration of 10 neurons in two consecutive layers to be the best performer. Across our experiments, ANNs were consistent, positioning it around the midpoint in terms of performance out of all the classifiers as seen by the relatively even colouring in Fig. 4. Interestingly however, denoising exhibited mixed effects, improving accuracy by 2–3% in the Tarcutta dataset, but decreasing it by approximately the same for Wambiana. Conversely, NB performed the least effectively among the studied classifiers. Regardless of the MFCC vector dimensionality, it repeatedly underperformed. Despite this, some FR, especially UMAP, bolstered its

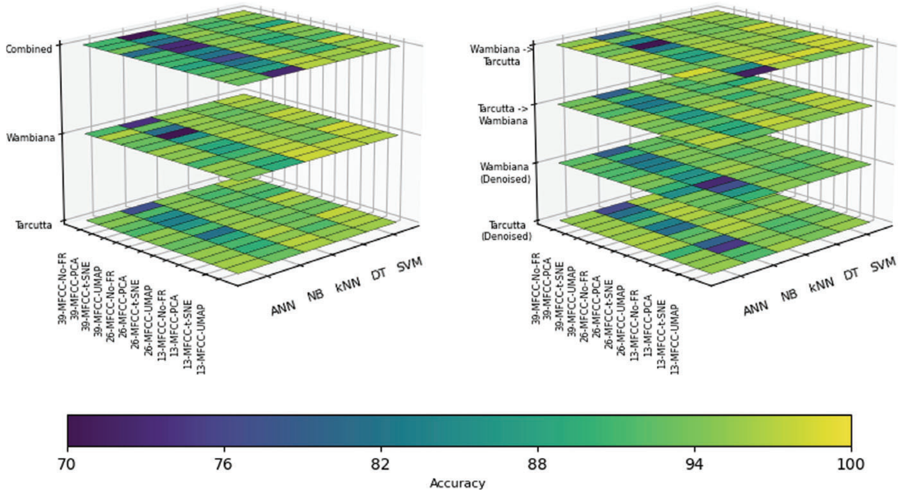
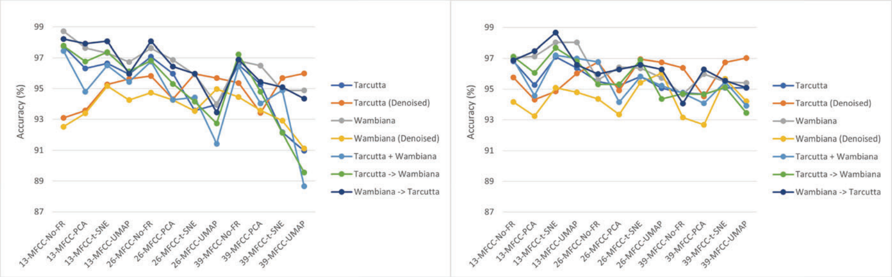


Fig. 4. Dual 3D stacked heatmaps showcasing the accuracy performance of each dataset on the Z-axis with respect to each classifier (ANN, NB, k-NN, DT, SVM) on the X-axis, MFCC feature vector (13, 26, 39) and FR combination (PCA, t-SNE, UMAP, No-FR) on the Y-axis.

capabilities, but it was still not able to elevate NB significantly enough. Similarly for DTs, although the 13-MFCC feature vector offered some increased performance when combined with t-SNE, it still lagged behind the leading classifiers. For the k-NN classifier algorithm we determined that the optimal k-value was $k = 5$, as it obtained the best classification performance across each dataset. With this, k-NN was able perform relatively well, particularly when paired with UMAP and t-SNE embeddings. However, SVM emerged as the strongest performer among the classifiers, achieving the highest accuracy in four out of the seven datasets used. Interestingly, the 13-dimensional MFCC vector, when combined with No-FR, maximised its class-separation capability. This is evidenced by 97.77%, 98.27%, 97.45% and 97.78% accuracy for the Tarcutta, Wambiana, their combination, and when training with Tarcutta and testing with Wambiana, respectively.

5.2 Discussion

In evaluating the multitude of algorithms employed in this study, SVM emerged as the standout performer in classifying the unique natural soundscape datasets, with k-NN in a close second. As seen in Fig. 5(a), while SVMs consistently achieved the highest accuracy, k-NN's effectiveness was more stable across the experiments as shown in Fig. 5(b). SVM's superior performance highlights its adaptability and efficiency in parsing the complexities inherent in such data. This goes against the findings of [6], who found ANNs with MFCCs to be as equally performant for their natural soundscape dataset. However, we have



(a) SVM Results.

(b) k-NN Results.

Fig. 5. Comparative line charts highlighting the effectiveness of the two best performing classifiers of (a) SVM and (b) k-NN against each dataset with respect to each MFCC feature vector (13, 26, 39) and FR combination (PCA, t-SNE, UMAP, No-FR).

shown through our experimental design that our approach has more proven flexibility across a range of ecoregions, as opposed to a dataset derived from a single study site. As for the feature vectors, Table 2 shows that classification results were consistently the highest when utilising 13 MFCCs. This goes against our initial assumptions and indicates that these coefficients alone are capable of capturing the underlying audio features without the need for additional temporal detail. The only exception to this case is when signal detail is lost through other means, such as denoising, as signified by the best performing methods for the denoised datasets using higher-MFCC feature vectors.

Table 2. The best classification performance for each dataset using the optimal method.

Dataset	Method	Accuracy	Precision	Recall	F1-score
Tarcutta	13-MFCC+No-FR+SVM	97.77	96.82	95.90	96.31
Wambiana	13-MFCC+No-FR+SVM	98.27	98.38	97.70	97.98
Tarcutta (Denoised)	39-MFCC+UMAP+k-NN	97.03	97.18	97.08	97.08
Wambiana (Denoised)	26-MFCC+UMAP+k-NN	96.04	95.73	95.05	94.90
Tarcutta + Wambiana	13-MFCC+No-FR+SVM	97.45	96.40	95.85	96.09
Tarcutta → Wambiana	13-MFCC+No-FR+SVM	97.78	96.54	96.02	96.27
Wambiana → Tarcutta	13-MFCC+t-SNE+k-NN	98.66	98.63	98.22	98.42

With respect to FR, the efficacy of t-SNE and UMAP in comparison to PCA is conditional based on the type of dataset and classifier used, as evidenced by the varying results in Fig. 5. It is of interest to note, that while SVM had better outcomes with No-FR, other classifiers, like k-NN, consistently benefited from this step. Furthermore, the classifiers demonstrated consistent performance across both the Wambiana and Tarcutta datasets. Regardless of the wide variance in sounds, or the disparity in dataset sizes (6,604 compared to 2,238 samples), the models maintained consistency, further demonstrating the flexibility of the proposed framework. Additionally, our cross-dataset validation experiments revealed not only great robustness in our approach, but also affirmed its potential for generalising across diverse ecoregions. Our optimised models were purposefully not conditioned on a particular set of species, nor were the datasets they were trained on overly sanitised. Instead, our models were trained with the nuances of real-world natural soundscapes included. While this has been achieved in other studies, such as [16], we have demonstrated a significant accuracy improvement upon this. Despite this, our approach was still able to achieve comparable results to studies which do use single-species datasets such as [1,3] and [8]. Conversely, while spectral subtraction as a noise reduction technique appeared promising initially, in practice, it removed too much signal information. From our experiments using the denoised versions of the Tarcutta and Wambiana datasets, models generally saw higher accuracy performance without spectral subtraction. This reaffirms the inherent challenges in this approach, as it can be difficult to generate a single noise profile to cover the wide variance of sounds across two distinct ecosystems and may be more suitable for datasets with a narrower focus [21].

6 Conclusion

Australia has some of the richest biodiversity globally, and it is imperative to monitor its species using sound, particularly those at risk. Until now, there has been a lack of transferability in model design across multi-ecoregion and multi-species datasets. Our methodological design, incorporating a range of supervised classifiers, pivoted around these challenges through the use of two distinct natural soundscape datasets. From this, we were able to conclude that for the majority of cases, 13 MFCCs, with No-FR applied, with SVM as the classifier, is consistently superior for this task. Moreover, this indicates that large-scale ecoacoustics datasets, transformed under this proposed framework, may be linearly separable in high-dimensional space and implies that, given a similar dataset, SVM may provide reliable classifications. Since we focused on incorporating the nuances of real-world natural soundscapes across different ecoregions, we believe that our framework is generalisable. Cross-dataset validation reinforces this, as high accuracy was maintained despite the large sound variance at each site. With this, we have strong supporting evidence that shows our proposed framework improves upon pre-existing approaches, is accurate and robust, and may serve as an ideal base for future ecoacoustics classification tasks.

References

1. Bardeli, R., et al.: Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. Pattern Recogn. Lett. **31**(12), 1524–1534 (2010)
2. Cardinale, B.J., et al.: Biodiversity loss and its impact on humanity. Nature **486**, 59–67 (2012)

3. Colonna, J., et al.: Automatic classification of anuran sounds using convolutional neural networks. In: Proceedings of the Ninth International C* Conference on Computer Science & Software Engineering, C3S2E 2016, pp. 73–78. Association for Computing Machinery, New York (2016)
4. Eichinski, P., et al.: A convolutional neural network bird species recognizer built from little data by iteratively training, detecting, and labeling. Front. Ecol. Evol. **10** (2022)
5. Gibb, R., Browning, E., Glover-Kapfer, P., Jones, K.E.: Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. Methods Ecol. Evol. **10**(2), 169–185 (2019)
6. Grinfeder, E., et al.: Soundscape dynamics of a cold protected forest: dominance of aircraft noise. Landscape Ecol. **37**(2), 567–582 (2022)
7. Krause, B.: Anatomy of the soundscape: evolving perspectives. J. Audio Eng. Soc. **56**(1/2), 73–80 (2008)
8. LeBien, J., et al.: A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network. Eco. Inform. **59**, 101113 (2020)
9. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(11) (2008)
10. McInnes, L., et al.: UMAP: uniform manifold approximation and projection. J. Open Source Softw. **3**(29), 861 (2018)
11. Mcloughlin, M.P., et al.: Automated bioacoustics: methods in ecology and conservation and their potential for animal welfare monitoring. J. R. Soc. Interface **16**(155), 20190225 (2019)
12. Mesaros, A., et al.: Detection and classification of acoustic scenes and events: outcome of the DCASE 2016 challenge. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(2), 379–393 (2018)
13. Pijanowski, B.C., et al.: Soundscape ecology: the science of sound in the landscape. Bioscience **61**(3), 203–216 (2011)
14. Roe, P., et al.: The Australian acoustic observatory. Methods Ecol. Evol. **12**, 1802–1808 (2021)
15. Salamon, J., et al.: Towards the automatic classification of avian flight calls for bioacoustic monitoring. PLoS ONE **11**(11), 1–26 (2016)
16. Scarpelli, M.D.A., et al.: Multi-index ecoacoustics analysis for terrestrial soundscapes: a new semi-automated approach using time-series motif discovery and random forest classification. Front. Ecol. Evol. **9**, 738537 (2021)
17. Stowell, D., et al.: Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge. Methods Ecol. Evol. **10**(3), 368–380 (2019)
18. Towsey, M., et al.: Long-duration, false-colour spectrograms for detecting species in large audio data-sets. J. Ecoacoustics **2**(1), 1–13 (2018)
19. Trawicki, M., Johnson, M., Osiejuk, T.: Automatic song-type classification and speaker identification of Norwegian ortolan bunting (Emberiza Hortulana) vocalizations. In: 2005 IEEE Workshop on ML for Signal Processing, pp. 277–282 (2005)
20. Wu, Z., Cao, Z.: Improved MFCC-based feature for robust speaker identification. Tsinghua Sci. Technol. **10**(2), 158–161 (2005)
21. Xie, J., Towsey, M., Zhang, J., Roe, P.: Adaptive frequency scaled wavelet packet decomposition for frog call classification. Eco. Inform. **32**, 134–144 (2016)