Audio Engineering Society

Conference Paper

Presented at the AES International Conference on
Machine Learning and Artificial Intelligence for Audio
2025 September 8–10, London, UK

From CNN to Reservoir Computing: A New Perspective on Acoustic Scene Classification

Blind Review

Correspondence should be addressed to ()

ABSTRACT

Acoustic Scene Classification (ASC) is a fundamental task in machine listening, aiming to categorize the acoustic environment in which an audio recording was made. Although Convolutional Neural Networks (CNNs) are still the dominant approach in ASC today, increasing attention has been given to optimizing model efficiency, particularly for model deployment in resource-limited computational platforms such as hearing aids. In this study, we explore Reservoir Computing (RC) as an alternative lightweight approach for ASC and systematically compare its performance against multiple CNN architectures, including ResNet-34, MobileNet-V2, EfficientNet-B2, and TF-SepNet-20. Our results indicate that CNN models consistently outperform RC in classification accuracy, even when trained without data augmentation. However, RC exhibits significantly lower computational cost, requires less training time, and achieves the fastest inference speed among all models. This trade-off between accuracy and efficiency suggests that model selection should be guided by application requirements: CNNs are preferable when accuracy is paramount, while RC provides an alternative for low-power real-time applications such as hearing aids.

1 Introduction

Hearing loss is a growing global health concern, affecting more than 11% of the population in the European Union, with projections that more than 2.5 billion people worldwide will suffer from hearing impairment by 2050 [1]. Despite the availability of hearing aids, only 37-41% of individuals with hearing difficulties use these devices, mainly due to persistent issues with speech comprehension in noisy environments and the complexity of manual adjustment processes [2, 3].

Addressing these challenges, modern hearing aid technology incorporates adaptive signal processing systems to enhance speech perception in complex acoustic environments [4]. A key objective is to enable hearing aids to dynamically adjust to individual hearing profiles of users and daily auditory scenes, improve speech intelligibility in noisy conditions, and adapt to diverse listening environments [3]. A fundamental functionality in achieving this adaptability is Acoustic Scene Classification (ASC), which involves recognizing and categorizing environmental soundscapes into predefined classes, such as shopping malls, metro stations, or urban streets [5]. Incorporating ASC into hearing aids allows real-time adaptation to changing acoustic scenes, improving speech intelligibility and user comfort. For example, an ASC module can detect a transition from a quiet office to a noisy street and automatically adjust the noise suppression and microphone directivity accordingly [3, 6].

However, the development of ASC models for hear-

ing aids introduces a critical trade-off between accuracy and computational efficiency. Hearing aids operate under strict resource constraints, requiring models that achieve robust scene recognition while minimizing memory usage and power consumption [7, 8]. Therefore, the design and optimization of learning architectures play a pivotal role in determining the overall performance of ASC systems. Convolutional Neural Networks (CNNs) have proven to be highly effective in developing ASC systems [9]. Given the increasing demand for efficient audio classification models, recent research has focused on exploiting the powerful feature extraction capabilities of CNNs to improve both the accuracy and computational efficiency of ASC systems [7]. However, while CNNs remain the most popular choice, recent research suggests that Reservoir Computing (RC) offers a lightweight alternative [10], especially suitable for neuromorphic sensor integration.

RC is an efficient framework for processing time-dependent data. It uses a fixed recurrent network (the reservoir) to project input signals into a high-dimensional space, training only the output layer. This approach significantly reduces training time and computational cost compared to traditional recurrent neural networks [10, 11]. The concept was developed to overcome the challenges of training complex neural networks using a fixed and randomly connected reservoir that captures rich temporal dynamics with minimal resource use [12, 13]. Here in this work, we want to exploit the dynamic response of the sensor for ASC in a RC setup as an energy efficient alternative to the neural networks models.

In ASC, the task is to identify the environment from an audio recording, which requires capturing both temporal and non-linear sound characteristics. RC is well suited for this purpose because it naturally processes sequences and highlights dynamic features such as amplitude changes, frequency shifts, and temporal patterns, improving the distinction between scenes such as busy streets, parks, and bustling airports [14]. Moreover, recent advances allow the reservoir to be implemented with physical devices such as lasers or memristors. These hardware reservoirs can perform the high-dimensional projection mechanically, further reducing computational costs. Optimizing hyperparameters directly on such devices offers a promising path toward faster, more energy-efficient, and potentially more accurate audio scene classification systems [15, 16, 17, 18, 19].

Selecting the optimal learning architecture is crucial for achieving accurate and efficient ASC performance in hearing aids. This study makes the following contributions: (1) we introduce a novel RC model for ASC, offering an alternative to conventional deep learning approaches; (2) we systematically compare the RC model with four CNN architectures—ResNet-34, EfficientNet-B2, MobileNet-V2, and TF-SepNet-20—evaluating their performance; (3) we provide a detailed analysis of the trade-offs between classification accuracy, computational complexity, and efficiency, highlighting how different architectures align with the constraints of real-world hearing aid applications.

2 Methods

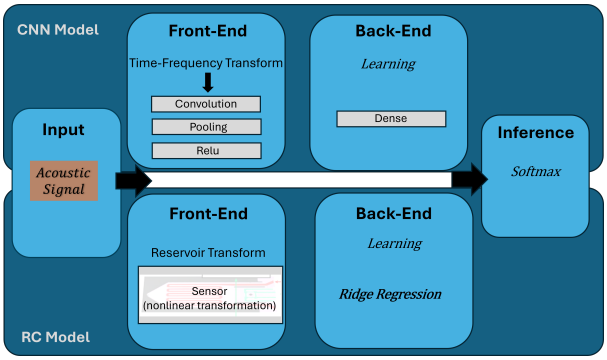

Fig. 1: Overview of the ASC processing pipeline.

To systematically compare CNN-based and RC-based approaches for ASC, we adopt a structured framework that decomposes each model into two key components: a front-end for feature extraction and a back-end for classification, as illustrated in Fig. 1.

In the front-end of an ASC model, raw audio input is transformed into a structured representation suitable for machine learning. In the case of common CNN-based models, this process involves applying the Short-Time Fourier Transform (STFT) followed by a Mel-scaled filter bank transformation and a logarithmic magnitude compression. Such Mel spectrograms are then processed by a series of convolutional layers, pooling operations, and non-linear activations, progressively extracting relevant acoustic features [20, 21]. Conversely, in the RC-based approach, the front-end consists of a reservoir transformation where the audio waveform is fed into a nonlinear dynamic system (the sensor). A Fast Fourier Transform (FFT) is then applied to extract

frequency-domain features, forming the representation used for classification.

The back-end of an ASC system is responsible for learning a mapping from the extracted features to predefined scene categories. For CNNs, this stage consists of fully-connected dense layers followed by a Softmax classifier, which assigns probabilities to different scene classes [6]. In contrast, the RC model employs ridge regression as a lightweight classification method, using the transformed reservoir responses for the final scene classification.

The following sections provide a detailed description of both CNN-based and RC-based ASC systems. Section 2.1 outlines the architecture and processing pipeline of the CNN approach, while Section 2.2 details the implementation of RC for ASC.

2.1 Convolutional Neural Network

CNNs are a class of deep learning models widely used for recognition tasks. Their ability to automatically extract spatial hierarchies of features from input data makes them particularly effective for image and audio classification tasks, including ASC [22, 23]. In the context of ASC, spectrogram representations are used as input features for CNNs to capture the spectral-temporal structures characteristic of different acoustic scenes.

Convolutional Layers. The convolutional layer is a core building block of CNNs. It extracts local features from the input by applying a set of learnable kernels (or filters) to the input feature map $\mathbf{X} \in R^{M \times N}$, resulting in a feature map $\mathbf{S} \in R^{M' \times N'}$. The convolution operation can be expressed mathematically as:

$$S_{i,j} = (\mathbf{K} * \mathbf{X})_{i,j} = \sum_{u=1}^{k} \sum_{v=1}^{k} K_{u,v} \cdot X_{i+u,j+v} \qquad (1)$$

where $\mathbf{K} \in R^{k \times k}$ is the convolutional kernel, and i, j denote the spatial position in the output feature map.

Several important parameters govern the behavior of the convolution operation, shaping the characteristics of the resulting feature maps. One such parameter is the **stride** (s_k), which determines the step size at which the kernel moves across the input [24]. Another parameter is **zero-padding** (p_k), which involves symmetrically adding rows and columns of zeros around the input to control the output size [25]. Additionally, the **dilation rate** (d_k) significantly influences the receptive field by introducing gaps between kernel elements [26].

Receptive Field. The receptive field of a neuron in a CNN refers to the region in the input space that contributes to its activation [27]. It determines the context in which a neuron can "see" at a specific layer.

Pooling Layers. Pooling layers downsample the feature maps, reducing their spatial dimensions while retaining essential information [28]. Two common pooling techniques are **Max Pooling** and **Average Pooling**. Max pooling retains the maximum value within each pooling window, emphasizing the most salient features [29]. Average pooling computes the average value within the pooling window, smoothing the feature map:

$$S_{i,j} = \frac{1}{k^2} \sum_{u=1}^{k} \sum_{v=1}^{k} X_{i+u,j+v} \qquad (2)$$

2.1.1 Feature Extraction for CNN

To effectively process audio inputs for our CNN models, as illustrated in Figure 1, we first apply a preprocessing pipeline to extract time-frequency representations. The input audio is provided in a single-channel format with a sampling rate of 44.1 kHz. Specifically, the STFT is performed using a window size of 3072 samples (64 ms) and a hop size of 500 samples (10.42 ms). Log-Mel spectrograms are computed using a Mel-scaled filter bank with 256 frequency bands and 4096 FFT points. Each spectrogram is stored as a three-dimensional tensor $\mathbf{X} \in R^{256 \times 84 \times 1}$, capturing 256 frequency bins, 84 time frames, and one channel which is processed by the CNN model.

2.1.2 CNN Architectures

As a performance comparison for the RC model in ASC, we evaluated four commonly used CNN architectures: ResNet-34 [30], EfficientNet-B2 [31], MobileNet-V2 [32], and TF-SepNet-20 [33].

ResNet-34 ResNet [30] is a family of deep CNNs designed to overcome the vanishing gradient problem through residual connections, which facilitate stable training in deep architectures. Among its variants, ResNet-34 adopts a bottleneck-free residual block design, preserving spatial and temporal details while maintaining a moderate parameter size. In this study, ResNet-34 is selected for its strong feature extraction capabilities and effective hierarchical modeling of spectro-temporal patterns, making it a well-suited candidate for ASC tasks [6, 34].

MobileNetV2 MobileNet [35] is a family of lightweight CNNs designed for efficiency, particularly in resource-constrained environments. It employs depthwise separable convolutions to reduce computational complexity, while maintaining strong feature extraction capabilities. Among its variants, MobileNetV2 [32] introduces inverted residual blocks, which enhance representational power with minimal computational cost. In this study, we adopt MobileNetV2 for its balance between accuracy and efficiency [36, 37].

EfficientNet-B2 EfficientNet [31] is a family of CNNs designed for optimal model scaling through compound scaling, which balances depth, width, and resolution to improve accuracy while minimizing computational costs. Among its variants, EfficientNet-B2 offers an trade-off between efficiency and performance, making it particularly suitable for ASC tasks that require spectro-temporal pattern recognition with low computational overhead. In this study, EfficientNet-B2 is selected for its lightweight architecture, swish activation for improved gradient flow, and mobile inverted bottleneck convolution (MBConv) layers for enhanced efficiency [37, 38].

TF-SepNet-20 TFSepNet [33] is an efficient CNN architecture designed to enhance feature extraction while minimizing computational cost. Unlike conventional CNNs, it employs time-frequency separable convolution (TF-SepConv), where 1D convolutional kernels process spectral and temporal features independently before merging them, optimizing ASC-related feature extraction. The architecture is inspired by BC-ResNet [39] but replaces 2D convolutions with 1D kernels along time and frequency axes, reducing parameter counts and MACs consumption. Additionally, depthwise separable convolutions and adaptive residual normalization improve efficiency and generalization. The global average pooling layer ensures a compact final representation while maintaining classification performance [6, 33].

2.1.3 Training Setup

The training process involves optimizing the model for 100 epochs using the Adam optimizer with default hyperparameters within the PyTorch framework. A batch size of 32 is used. To facilitate faster convergence and to improve training stability, a warm-up learning rate strategy is adopted [40]. Specifically, the learning rate is gradually increased from 0 to 0.01 over the initial five epochs. After the warm-up phase, the learning rate is progressively reduced to 0 following a cosine annealing schedule [41]. Furthermore, early stopping is used to prevent overfitting, using 20% of the training data as validation data, with a patience setting of 15 epochs. Training is stopped if validation performance does not improve for 15 consecutive epochs [42]. Moreover, it is important to note that the CNN architectures in this study were trained without data augmentation.

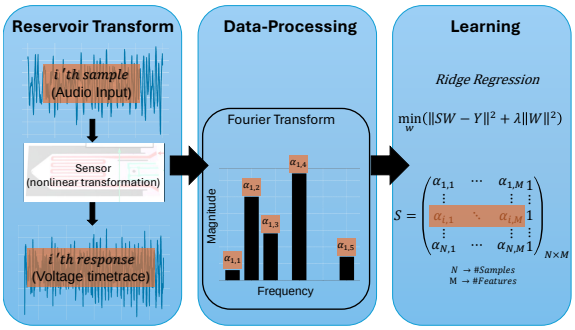

Fig. 2: The overall pipeline for the sensor based RC method using a MEMS-sensor for ASC.

2.2 Reservoir Computing Model Design

The overall pipeline of the RC method is shown in Figure 2. In the following we explain each component in detail.

Sensor-Reservoir Transform. We simulate the dynamics of a microelectromechanical (MEMS) sensor using equations derived from the Euler-Bernoulli beam theory, as detailed in [15]. The MEMS sensor is described by the following set of equations:

$$\ddot{x}(t) + \frac{\omega_0}{Q_0}\dot{x}(t) + \omega_0^2 x(t) = \alpha\,\theta(t) + F_{\text{ext}}(t) \tag{3}$$

$$\dot{\theta}(t) + \beta\,\theta(t) = \frac{\gamma \cdot \min((a\,u_{\text{ac}}(t) + u_{\text{dc}})^2, u_{\text{max}}^2)}{R^2} \tag{4}$$

$$\dot{u}_{\text{ac}}(t) = -\frac{u_{\text{ac}}(t)}{\tau} + k\dot{x}(t) \tag{5}$$

Here $x(t)$, $\theta(t)$, and $u_{ac}(t)$ are beam deflection, temperature change and actuation voltage and describe the time dependent oscillation state of the sensor. The sensor is designed to mimic the behavior of the cochlea by converting external forces (i.e., sounds modelled by F_{ext}) into voltage time series $u_{ac}(t)$ using piezoresistive elements and an in-time feedback loop (feedback

strenth a in Eq.(4)) for improved adaptation to its surroundings. The parameters $\alpha, \omega_0, Q_0, \gamma, \beta, \tau, k, R, u_{dc}$, $u_{\max}$ are conversion factor (deflection $x(t)$ to sensing voltages), angular resonance frequency, quality factor, conversion factor (actuation voltage u_{ac} to temperature θ), time constant (temperature changes), time constant for high pass filter, calibration factor (deflection $x(t)$ to sensing voltage), heater resistance, bias voltage, and cutoff voltage, respectively, and chosen to model the actual device from [15].

We implement the parameters for an existing sensor [15] and derive guideline for experimental implementation within a RC setup. We focus on the effect of the feedback strength a and the dc voltage u_{dc}.

To fully understand the sensor's behavior, its dynamics are first analyzed in the absence of external forces. This analysis includes investigating resonance frequencies, bifurcations, and overall sensor responses, which are crucial for effective hyperparameter optimization. It has already been shown that the sensor dynamics exhibits a Hopf bifurcation in (u_{dc},a)-parameter space, beyond which self sustained oscillations occur[43]. In Fig.3a, the steady state amplitude x of the sensor response is color coded with grey while the number of maxima observed in x in the self oscillatory regimes is depicted with blue/pink colors. The Hopf bifurcation is visualized via white dashed lines as determined numerically from our simulations. The yellow dashed line represents the dynamic transition from steady state to the self-oscillatory low-frequency dynamics which is due to the clipping factor in Eq. (4). Exemplary time series of the sensor deflection $x(t)$ are shown in Fig.3b,c.

Comparing positive and negative feedback strength, we find that the sensor exhibits two distinct dynamic regimes which are: 1) A regime showing high resonance frequency near 14 kHz for negative a (timeseries in Fig. 3c), and 2) A regime with a complex dynamic response consisting of low frequencies (from 40 Hz to 250 Hz) and the 14 kHz component (timeseries in Fig. 3b).

Simulation and Data-Processing. We simulate the RC performance of the sensor using an audio scene dataset composed of one-second mono recordings at a 44,100 Hz sample rate. These recordings serve as the external force $F_{ext}(t)$ that acts on the sensor. The dynamic response of the sensor is given by the ODE

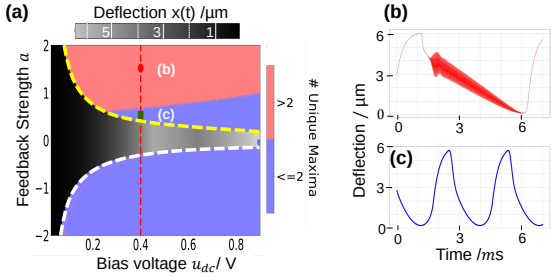

Fig. 3: Sensor dynamics (a) Unique maxima found in the sensor deflection $x(t)$ color coded in blue/red as a function of bias voltage u_{dc} and feedback strength a. Grey colors depict deflection value for the case without oscillations. White dashed line represents the Hopf-bifurcation and yellow dashed line represents the dynamic transition induced by the min-function in Eq. (4). (b) Time series of $x(t)$ with both high and low frequency components at $u_{\mathrm{dc}} = 0.4$, $a = 0.52$, (c) $x(t)$ with only low-frequency components at $u_{\mathrm{dc}} = 0.4$, $a = 1.52$.

system presented in Eqs.(3)-(5) which we simulate using a fourth-order Runge-Kutta method with a time step of 10^{-6}. The audio samples are first interpolated, passed through a low-pass filter to eliminate aliasing effects, and then fed into the ODE system in order to obtain the sensor response u_{ac}. Subsequently we post-process the response u_{ac} by applying a Fast Fourier Transform (FFT) to time series where each has a length of 1×10^6. Each sequence i of a Fourier transformed time series is one sample (number of samples is N). The FFT is given by

$$X(k) = \sum_{n=0}^{N'-1} x(n)\, e^{-j\frac{2\pi kn}{N'}}, \quad k = 0, 1, \ldots, N'-1, \quad (6)$$

where $N' = 10^6$ is the length of each time series and n is the time index (the time between n and n+1 ,which is one microsecond). We then compute the magnitude $|X(k)|$ via $|X|^2 = (\mathrm{Re}X)^2 + (\mathrm{Im}X)^2$ for each frequency k, where each k represents one feature of the sensor response. The total number of features M will be subject to change as we can select different numbers of relevant features for classification.

For the classification task, we select M frequencies (features) that are most informative for distinguishing

between classes. The M magnitudes $|X(k)|$ (denoted as α_i for sample i in Fig.2) form the feature vector for each sample. The resulting state matrix S has M columns (features) and N rows (samples).

Learning via Ridge Regression. Once the state matrix S is filled with the simulated and Fourier transformed sensor responses, we normalize the obtained data so that each feature (each column in S) has a zero mean and unit variance. We then append an additional column of ones to the state matrix S to serve as a bias term[44]. We use RC for the training and thus we will need 10 ridge regression with Tikhonov regularization (parameter λ) to classify the data into ten distinct classes. The trained weights that determine the output of the reservoir computer are called weight matrix W and are determined from solving the optimization problem $\min_W \ \|Y - SW\|_2^2 + \lambda \|W\|_2^2$ with Y being the target vector. The closed-form solution is given by

$$\hat{W} = \left(S^T S + \lambda I\right)^{-1} S^T Y. \tag{7}$$

A winner-takes-all strategy is then applied to determine the final predicted class for each sample.

Hyperparameter Optimization. In the RC approach, two key groups of hyperparameters are optimized: the sensor parameters which are bias voltage u_{dc} and feedback strength a, and the model parameters which are ridge regression regularization parameter λ and the number of features M. In order to find the optimal sensor parameters u_{dc} and a, we perform a grid search. Fig. 4 shows the performance values achieved when scanning the u_{dc}-a-parameter plane. Within the steady state regions (grey colors in Fig.3) we find the best performance (red colors in Fig.4) . Also, slight improvements are seen when the operation point is closer to the transition point from steady state to low-frequency regime (yellow line in Fig.4).

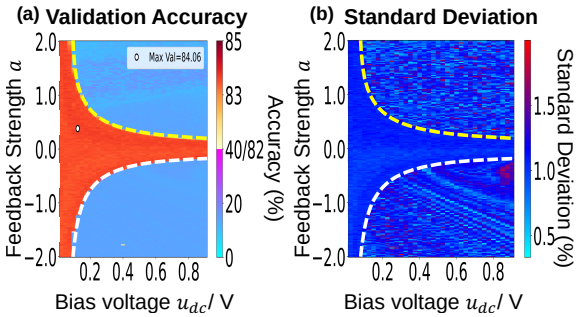

Fig. 4: Sensor Parameters (a) Average validation accuracy (from a 10-fold cross-validation) of the RC model as a function of sensor settings—bias voltage u_{dc} and feedback strength a. The black dashed line marks the Hopf bifurcation, and the model uses $M = 24000$ features. (b) The corresponding standard deviation of these accuracies; note that the high-accuracy (red) region in (a) has a standard deviation of 1.

3 Evaluation and Result

3.1 Dataset

In this work, we use the TAU Urban Acoustic Scenes 2022 Mobile development dataset [14], which is a well-established benchmark to assess the performance of low-complexity ASC systems. The dataset contains audio recordings from four different devices. However, for our experiments, we exclusively use the recordings from device A. This device captures the highest-quality audio using a Soundman OKM II Klassik/studio A3 binaural microphone and a Zoom F8 recorder. The decision to limit our analysis to a single device was made to minimize variability caused by different recording setups, ensuring that any observed performance gains are driven solely by the proposed model architectures. This approach allows us to isolate the impact of the model design on classification accuracy without introducing domain shifts or device-specific effects. We follow the official 7:3 training/test split for our experiments.

3.2 Experimental Design

To systematically evaluate the performance of the RC model in comparison with CNN architectures under different data conditions, we design two experiments. The first experiment investigates model performance under limited data diversity by restricting the training set to a single city, specifically Barcelona, from the dataset. The training data is split into 80% for training and 20% for validation, and the same split is applied across all models to ensure a fair comparison. This setting allows us to assess the performance of different models when data is scarce.

The second experiment examines the effect of increasing data diversity. We expand the training data to include recordings from three and five different cities, enabling us to analyze how increasing the number of

cities in the training data impacts the performance of the RC model compared to CNN architectures. Both experiments are conducted using the four CNN models ResNet-34, EfficientNet-B2, MobileNetV2, and TF-SepNet, as described in Section 2.1.2. This setup allows us to explore the relationship between model size and classification ability across different data conditions and to better understand whether the RC model is particularly advantageous when training data is limited.

3.3 Computational Complexity

When comparing the computational complexity of RC and CNN models, it is important to acknowledge that the methods for calculating the number of parameters and Multiply-Accumulate operations (MACs) differ significantly due to their distinct architectural structures.

Parameter Calculation in CNNs. In CNN, the total parameter count is obtained by summing up the parameters from all convolutional and fully connected layers [45].

MACs Calculation in CNNs. The total MACs is defined as the sum of preprocessing MACs and CNN MACs. The preprocessing MACs account for the operations required to transform raw audio waveforms into log-Mel spectrograms, which serve as inputs to the CNN. Moreover, to facilitate the calculation of MACs in our CNN models, we employed the open-source tool *NeSsi* [45].

Parameter Calculation in RC. In the ridge regression classifier for a 10-class problem in RC for this work, the state matrix S is composed of N rows (samples) and M columns (features). To account for the bias term, an additional column of ones is appended to S, resulting in an effective feature dimension of $M+1$. Consequently, the weight matrix W that maps the augmented feature space to the 10 output classes has dimensions $(M+1)\times 10$. Therefore, the total number of parameters in the model is given by $(M+1)\times 10$.

MACs Calculation in RC. The total MACs in RC is also defined as the sum of data-processing MACs and RC model MACs. The data-processing MACs account for the operations required to apply the real FFT on sensor responses [46]. Moreover, in order to have similar measure as the CNN models, the calculation of MACs in RC model also employed the open-source tool *NeSsi* [45].

3.4 Results

The results of **Experiment 1** are summarized in Table 1. Regarding validation accuracy, all three CNN architectures exhibit consistently high performance, with values exceeding 95%. In contrast, the RC model shows a lower validation accuracy of 84.49%. In terms of test accuracy, the CNN models also demonstrate stable performance, with all three basic architectures achieving results above 60%. TF-SepNet-20, while having a lower test accuracy than the other CNNs, still achieves 56.00%. The RC model, on the other hand, reaches a test accuracy of 44.50%. Regarding model complexity, both TF-SepNet-20 (0.016M) and the RC model (0.012M) exhibit significantly lower parameter counts compared to the other three CNN architectures. In terms of training time (TT), TF-SepNet-20 is remarkably faster than the other CNN models, requiring only 557.89s. However, the RC model is the fastest in training, with an extremely low TT of just 0.21s. Moreover, the RC model achieves the fastest inference time (IT) overall, with a value of only 1.70 ms.

The results of **Experiment 2**, illustrated in Figure 5. For validation accuracy, the 4 CNN models show consistently high performance, maintaining values around 90% or higher under different training data conditions. In contrast, the RC model shows a noticeable decline in validation accuracy as the diversity of training data increases. Starting from 84.49% with 1 city, its validation accuracy drops to 72.38% with 3 cities and further decreases to 66.06% with 5 cities. In terms of test accuracy, all four CNN models remain relatively stable, fluctuating around 60%, with minor variations across different training conditions. The RC model also exhibits relatively stable test accuracy, maintaining values around 45%. However, unlike its validation accuracy, the RC test accuracy shows a slight upward trend as the diversity of training data increases, improving from 44.50% with 1 city to 49.75% with 5 cities. Regarding TT, the RC model demonstrates significantly lower computational requirements, with training times below 1s across all data conditions, highlighting its exceptional computational efficiency.

4 Conclusion

The experimental results highlight the trade-off between classification performance and computational efficiency in ASC models. The CNN models consistently outperformed RC in classification accuracy, with

Architecture	Params (M)	MACs (G)	TT (s)	IT (ms)	Valid. Acc. (%)	Test Acc. (%)
ResNet-34	21.283	29.81	266.44	3.93	95.39	62.75
EfficientNet-B2	9.214	5.82	993.13	10.87	98.10	64.75
MobileNet-V2	2.236	2.54	446.67	7.15	97.19	64.50
TF-SepNet-20	0.016	0.27	557.89	4.95	93.96	56.00
Reservoir Computing	0.012	0.01	0.21	1.70	84.49	44.50

Table 1: Experiment 1: Model Performance on a Single City Dataset. The table presents a comparative analysis of different architectures in terms of model complexity (Params, MACs), computational efficiency, and classification accuracy (Validation Accuracy and Test Accuracy). TT refers to Training Time, and IT refers to Inference Time.

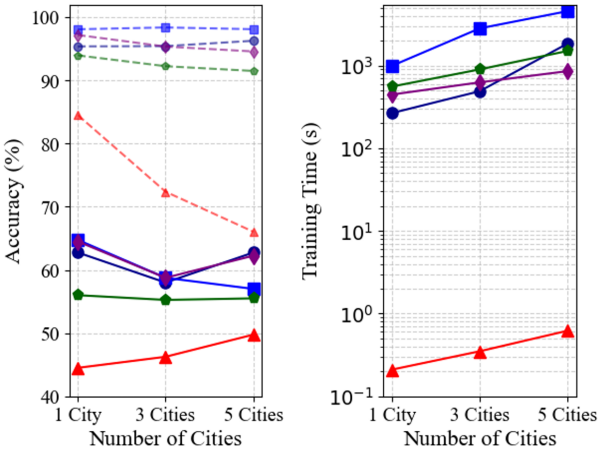

Fig. 5: Experiment 2: Impact of Increasing Data Diversity on Model Performance. Left: Test accuracy (%) and validation accuracy (%) trends across different city data splits. Solid lines represent test accuracy, while dashed lines indicate validation accuracy. **Right:** Training time across different models, displayed on a semi-logarithmic scale. The colors indicate different architectures: ResNet-34, EfficientNet-B2, MobileNet-V2, TF-SepNet-20, and Reservoir Computing.

all four CNN variants achieving test accuracy around 60%, while the RC model maintaining test accuracy around 45–50%. However, the RC model demonstrated remarkable efficiency in both TT and IT. Unlike CNNs, which require extensive training, the RC model required less than 1s for training across all experimental settings. Its IT remained significantly lower than that of CNNs, making it an attractive option for real-time applications with strict computational constraints.

Rather than presenting a best model, our findings suggest two viable solutions depending on system requirements: CNNs for accuracy-driven applications and RC for efficiency-critical scenarios. This flexibility can be beneficial, as it allows system designers to tailor ASC implementations according to specific needs. The ability to select between a high-accuracy model and an ultra-efficient model provides a pathway for optimizing ASC deployment in practical applications, particularly for hearing assistive technologies where balancing accuracy and computational cost is crucial.

This work presents an initial and straightforward step towards developing an RC model for ASC. While our results are promising, there is significant potential for further improvement. Future research may focus on optimizing the sensor setup and incorporating multiple sensors in an optimal configuration. Such advancements could lead to enhanced performance, potentially allowing the system to operate directly on the raw reservoir time series responses without the need for extensive intermediate data processing.

References

[1] Organization, W. H., "World Report on Hearing," *World Report on Hearing*,

[2] Kochkin, S., "MarkeTrak VII: Obstacles to adult non-user adoption of hearing aids," *The Hearing Journal*, 60, p. 24–51,

[3] Vanitha Devi, R. and Vasundhara, "Review on Recent Advances in Hearing Aids: A Signal Processing Perspective," in R. P. Yadav, S. J. Nanda, P. S. Rana, and M. H. Lim, editors, *Proceedings of the International*

Conference on Paradigms of Computing, Communication and Data Sciences, Algorithms for Intelligent Systems, Springer, Singapore,

[4] Lenk, C., Ved, K., Durstewitz, S., Ivanov, T., Ziegler, M., and Hövel, P., "Bio-inspired, Neuromorphic Acoustic Sensing," in M. Ziegler, T. Mussenbrock, and H. Kohlstedt, editors, *Bio-Inspired Information Pathways*, volume 16 of *Springer Series on Bio- and Neurosystems*, Springer, Cham,

[5] Barchiesi, D., Giannoulis, D., Stowell, D., and Plumbley, M. D., "Acoustic Scene Classification: Classifying Environments from the Sounds They Produce," *IEEE Signal Processing Magazine*, 32(3), pp. 16–34,

[6] Ding, B., Zhang, T., Wang, C., Liu, G., Liang, J., Hu, R., Wu, Y., and Guo, D., "Acoustic Scene Classification: A Comprehensive Survey," *Expert Systems with Applications*, 238, p. 121902, 2024,

[7] Li, Y., Cao, W., Xie, W., Huang, Q., Pang, W., and He, Q., "Low-Complexity Acoustic Scene Classification Using Data Augmentation and Lightweight ResNet," in *2022 16th IEEE International Conference on Signal Processing (ICSP)*, volume 1, pp. 41–45,

[8] Sze, V., Chen, Y.-H., Yang, T.-J., and Emer, J. S., "Efficient Processing of Deep Neural Networks: A Tutorial and Survey," *Proceedings of the IEEE*, 105(12), pp. 2295–2329,

[9] Piczak, K. J., "Environmental Sound Classification with Convolutional Neural Networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6,

[10] Maass, W., Natschläger, T., and Markram, H., "Real-Time Computing without Stable States: A New Framework for Neural Computation Based on Perturbations," *Neural. Comput.*, 14(11), p. 2531, 2002,

[11] Jaeger, H., "The "echo state" approach to analysing and training recurrent neural networks," GMD Report 148, GMD - German National Research Institute for Computer Science,

[12] Larger, L., Baylón-Fuentes, A., Martinenghi, R., Udaltsov, V. S., Chembo, Y. K., and Jacquot, M., "High-Speed Photonic Reservoir Computing Using a Time-Delay-Based Architecture: Million Words per Second Classification," *Phys. Rev. X*, 7, p. 011015,

[13] Nakajima, M., Tanaka, K., and Hashimoto, T., "Scalable reservoir computing on coherent linear photonic processor," *Commun. Phys.*, 4(20), p. 2399,

[14] Heittola, T., Mesaros, A., and Virtanen, T., "TAU Urban Acoustic Scenes 2022 Mobile, Evaluation Dataset,"

[15] Lenk, C., Hövel, P., Ved, K., Durstewitz, S., Meurer, T., Fritsch, T., Männchen, A., Küller, J., Beer, D., Ivanov, T., and Ziegler, M., "Neuromorphic acoustic sensing using an adaptive microelectromechanical cochlea with integrated feedback," *Nat. Electron.*, 6(5), pp. 370–380, 2023,

[16] Antonik, P., Duport, F., Hermans, M., Smerieri, A., Haelterman, M., and Massar, S., "Online Training of an Opto-Electronic Reservoir Computer Applied to Real-Time Channel Equalization," *IEEE Trans. Neural Netw. Learn. Syst.*, 28(11), p. 2686,

[17] Larger, L., Soriano, M. C., Brunner, D., Appeltant, L., Gutierrez, J. M., Pesquera, L., Mirasso, C. R., and Fischer, I., "Photonic information processing beyond Turing: an optoelectronic implementation of reservoir computing," *Opt. Express*, 20(3), pp. 3241–3249,

[18] Vandoorne, K., Mechet, P., van Vaerenbergh, T., Fiers, M., Morthier, G., Verstraeten, D., Schrauwen, B., Dambre, J., and Bienstman, P., "Experimental Demonstration of Reservoir Computing on a Silicon Photonics Chip," *Nat. Commun.*, 5(1), p. 3541, 2014,

[19] Barazani, B., Dion, G., Morissette, J.-F., Beaudoin, L., and Sylvestre, J., "Microfabricated Neuroaccelerometer: Integrating Sensing and Reservoir Computing in MEMS," *J. Microelectromech. Syst.*, 29(3), pp. 338–347,

[20] Singh, V. K., Sharma, K., and Sur, S. N., "A survey on preprocessing and classification techniques for acoustic scene," *Expert Systems with Applications*, 229, p. 120520,

[21] Abesser, J., "A Review of Deep Learning Based Methods for Acoustic Scene Classification," *Applied Sciences*, 10(6), p. 2020,

[22] LeCun, Y., Bengio, Y., and Hinton, G., "Deep Learning," *Nature*, 521(7553), pp. 436–444,

[23] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, 60(6), p. 84–90, 2017,

[24] Dumoulin, V. and Visin, F., "A guide to convolution arithmetic for deep learning," *arXiv preprint arXiv:1603.07285*,

[25] Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., and Chen, T., "Recent advances in convolutional neural networks," *Pattern Recognition*, 77, pp. 354–377, 2018,

[26] Yu, F. and Koltun, V., "Multi-Scale Context Aggregation by Dilated Convolutions," *CoRR*, abs/1511.07122,

[27] Luo, W., Li, Y., Urtasun, R., and Zemel, R. S., "Understanding the Effective Receptive Field in Deep Convolutional Neural Networks," *ArXiv*, abs/1701.04128,

[28] Lin, M., Chen, Q., and Yan, S., "Network In Network," *CoRR*, abs/1312.4400,

[29] Zeiler, M. D. and Fergus, R., "Visualizing and Understanding Convolutional Networks," in D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, volume 8689 of *Lecture Notes in Computer Science*, pp. 818–833, Springer, Cham,

[30] He, K., Zhang, X., Ren, S., and Sun, J., "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778,

[31] Tan, M. and Le, Q. V., "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *ArXiv*, abs/1905.11946,

[32] Sandler, M., Howard, A. G., Zhu, M., Zhmoginov, A., and Chen, L.-C., "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520,

[33] Cai, Y., Zhang, P., and Li, S., "TF-SepNet: An Efficient 1D Kernel Design in Cnns for Low-Complexity Acoustic Scene Classification," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 821–825,

[34] 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," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 131–135,

[35] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *ArXiv*, abs/1704.04861,

[36] Howard, A. G., 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 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1314–1324,

[37] Zhang, X., Xie, W., and Gong, Y., "Lightweight Acoustic Scene Classification Using Knowledge Distillation and Efficient CNNs," *IEEE Transactions on Audio, Speech, and Language Processing*, 30, pp. 2055–2067,

[38] Ananya, I. J., Suad, S., Choudhury, S. H., and Khan, M. A., "A Comparative Study on Approaches to Acoustic Scene Classification Using CNNs," *ArXiv*, abs/2204.12177,

[39] Kim, B., Chang, S., Lee, J., and Sung, D., "Broadcasted Residual Learning for Efficient Keyword Spotting," in *Interspeech*,

[40] Goyal, P., Dollár, P., Girshick, R. B., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K., "Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour," *ArXiv*, abs/1706.02677,

[41] Loshchilov, I. and Hutter, F., "SGDR: Stochastic Gradient Descent with Warm Restarts," *arXiv: Learning*,

[42] Prechelt, L., "Early Stopping-But When?" in *Neural Networks*,

[43] Strogatz, S. H., *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*, CRC Press, Boca Raton, 2nd edition edition, 2015, , eBook Published 23 May 2019

[44] Hülser, T., Köster, F., Jaurigue, L. C., and Lüdge, K., "Role of delay-times in delay-based Photonic Reservoir Computing," *Opt. Mater. Express*, 12(3), p. 1214,

[45] Ancilotto, A., "NeSsi: Neural Network Simple Statistics Interface," `https://github.com/AlbertoAncilotto/NeSsi`, , accessed: 2025-01-31

[46] Cooley, J. W. and Tukey, J. W., "An algorithm for the machine calculation of complex Fourier series," *Mathematics of computation*, 19(90), pp. 297–301,