# Efficient Sound Classification Model using Raw Audio Waveforms in Resource Constrained Devices: Design, Implementation, and Evaluation

***Abstract*—Sound classification model *DPNet*, the paper describes a privacy-preserving deep learning model that is very light and can work on microcontrollers with limited resources for real-time bathroom sound classification. DPNet provides a very efficient approach by applying a compact Temporal Feature Extraction Block (TFEB) which is optimized for embedded inference, in contrast to the high computational cost existing models or privacy concerns raised by camera-based systems. The model is evaluated for real-time on an ESP32-S3 Nano (512 KB SRAM, 8 MB PSRAM). DPNet scores 99.21% on high-performance processors and 73.44% in real-time deployment on the dataset of seven bathroom activities created in-house. In comparison, ACDNet gets 99.64% and 62.31%, while ACLNet has 98.46% and 22.64% under similar conditions. DPNet, with a model size of only 4.5 MB, shows excellent accuracy–efficiency trade-offs; thus, it can be regarded as suitable for the privacy-aware and low-power acoustic monitoring.**

***Index Terms*—Sound Classification, Raw Audio Waveform, TinyML, Proposed Light-weight DPNet, Privacy-Preserving Sensing**

## I. Introduction

As automation and assistive technologies are increasingly adopted, the requirement for intelligent systems capable of recognizing human activities via *privacy-preserving* is growing. Among the different detection methods, *acoustic sensing* is considered a practical option for home and assistive applications [1], [2]. While cameras limit their use in environments where the users' privacy should be protected, microphones can still capture the non-verbal sounds happening around without disclosing anyone's identity thus becoming the best solution for places like bathrooms [2].

Despite all this, the challenge of recognizing bathroom sounds is still there and quite severe with overlap in frequency components, background noise and reverberation [3], [4]. At the same time, deep learning models traditionally excel in such controlled environments but their performance in terms of accuracy drops drastically in noisy and resource-starved settings [5], [6]. Furthermore, running big models on low-power microcontrollers is limited due to memory and computation restrictions, thus indicating the necessity of designing lightweight, energy-efficient architectures that are compatible with TinyML platforms [7]–[9].

This paper therefore introduces a noise-resilient and lightweight framework for real-time bathroom sound classification, which works on an **ESP32-S3 Nano** microcontroller. The system performs inference completely on-device, thus making it low power consuming and user privacy-protecting, which is suitable for those smart environments that are using assistive technology and need continuous monitoring.

Sound classification systems that are currently available quite often rely on cloud computation or high-performance processors, and they can't deal with the real-world acoustic challenges of echoes, cross-talk, and overlapping events. That is why there is an urgent requirement for an efficient model having strong on-device performance in such situations.

The major achievements of this research can be summarized as:

- **Lightweight Audio Classifier:** A small CNN design, *DPNet*, is revealed to distinguish seven different noise activities in a bathroom through a Temporal Feature Extraction Block (TFEB), hence, decreasing the number of parameters without losing the accuracy.
- **TinyML Deployment:** The DPNet model that has been quantized (int8–float32) is equipped on an *ESP32-S3 Nano* microcontroller with a PDM microphone and display unit, thereby facilitating efficient on-device inference under strict hardware constraints.
- **Real-Time Evaluation:** DPNet scores **73.44%** accuracy for real-time TinyML deployment, thus surpassing the existing models—ACDNet (62.31%) and ACLNet (22.64%)—not only in accuracy but also in compactness and energy efficiency.

The schedule of the rest of the paper is as follows: Section II presents a review of the related work; Section III is dedicated to the system prototype; Section IV describes the process of dataset preparation; Section V elaborates on the proposed model; Section VI reports on the evaluation results and the analysis of TinyML deployment; and Section VII wraps up the discussion with the conclusion of the paper.

## II. Related Work

The latest developments in TinyML have allowed application of deep learning inference even on ultra-low-power microcontrollers, thus, raising the bar for real-time acoustic classification on embedded devices. Previous literature on bathroom acoustics [2], [5], [6] employed Mel-Frequency Cepstral Coefficients (MFCCs) in combination with Hidden Markov Models (HMMs) for the identification of activities, such as flushing, showering, and brushing. Nevertheless, the proposed methods could not be evaluated positively in terms of robustness or reliability and they required manual calibra-

tion; plus, they were not suitable for real-time or large-scale applications.

Next, the research changed the model architecture from standard CNNs to lightweight CNNs to obtain better recognition results while keeping the cost of computation very low. Similar to [7], [10], [11], researchers got 85–97% accuracy with less than 250 KB models on microcontrollers such as Arduino Nano 33 BLE Sense for detecting keywords and voices. Studies in [12]–[14] increased the size of CNNs to recognize environmental and biomedical sounds from Mel-spectrogram or MFCC representations, which made real-time inference possible on memory-limited edge devices as the process was highly efficient.

Sound classification in bathrooms and hygiene-related areas performed by models such as ResNet, W-YAMNet, and MobileNet [1], [3], [4] exceeded the 90% accuracy mark in perfect conditions. Nevertheless, their dependence on high-performance hardware (like Raspberry Pi, K210, etc.) led to the inability of being deployed on smaller microcontrollers, and their performance dropped in noisy and reverberant conditions.

The recently introduced ACDNet [9], a compact CNN, can learn from the raw waveforms and apply it to the datasets such as ESC-50 and even run on STM32 microcontrollers. Other papers [8], [15] have looked into energy-efficient and neuromorphic architectures, and have stressed the point of hardware-aware model design for TinyML sound processing.

Nevertheless, even with the progress made in this area, the acoustic recognition system based on TinyML is still facing problems such as being too sensitive to noise, having small labeled datasets, and lacking portability across different real-world situations. The case of bathroom sound entropy and recognition on-resource-constrained microcontrollers has been hardly studied, which is the reason why the development of *DPNet*, a compact, noise-resilient, and privacy-preserving framework, is encouraged.

## III. Prototype Development

The system for bathroom sound classification is put into practice on a small embedded platform utilizing the **ESP32-S3 Nano** microcontroller as its core. The ESP32-S3, a product of Espressif Systems, comprises a dual-core 32-bit Xtensa LX7 processor, which can run at a maximum frequency of 240 MHz, along with 512 KB of on-chip SRAM and 8 MB of external PSRAM. Such a setup guarantees the availability of memory and processing power, sufficient for the real-time execution of the quantized neural network models.

The prototype in the picture (Fig. 1) consists of a Pulse Density Modulation (PDM) microphone and an OLED display unit. The PDM microphone picks up the sounds of the environment, like the sound of flushing, running water in the shower, and running tap water, which are completely processed in the ESP32-S3 Nano. Audio capturing and preprocessing are done through the $I^2S$ interface at a sampling rate of 20 kHz, then followed by real-time predictions with the help of the quantized TensorFlow Lite (TFLite) model.

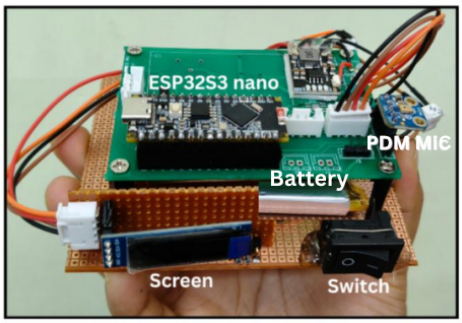


Fig. 1: Prototype hardware setup showing the ESP32-S3 Nano microcontroller, PDM microphone, display, battery and switch

## IV. Dataset Creation

As a part of the initiative to enable sound classification for bathrooms while keeping the privacy intact, an exclusive dataset containing seven sound classes was created: `Flush`, `No Class`, `Shower`, `Bathroom Tap`, `Basin Tap`, `Door`, and `Walker/Crutch` (for seniors or disabled users). The classes depict typical acoustic events occurring in residential bathroom settings.

The audio data were obtained with a microphone attached to a mobile device (see Fig. 1). The unprocessed recordings went through a process of manual annotation and segmentation where class labels were applied. Approximately 55 minutes of audio per class was the approximate duration of audio for each class, which made sure that the classes were represented equally.

In order to build a more robust model and facilitate better generalization, the technique of amplitude-based data augmentation [16] was utilized by adjusting each clip's volume ±25%. For each original sample, two augmented versions were created, thereby effectively increasing the total data from 55 to 165 minutes for each category. The recording sessions were held in 11 different places to ensure that different sound qualities and volume levels were used thus making it easier for the system to recognize sounds in real-time.

## V. Methodology

### A. Proposed DPNet Architecture

The proposed **DPNet** (Figure. 2) model is a small *real-time* bathroom sound classification convolutional neural network suitable for resource-constrained hardware like the ESP32-S3 Nano. It is inspired by the ACDNet [9], EnvNet-v2 [17], and ACLNet [18] methods, but importantly, it does so by providing key architectural simplifications that lead to less parameter count and computation while high accuracy is achieved.

DPNet uses a **Spectral Feature Extraction Block (SFEB)** similar to ACDNet to get low-level spectral cues from the human auditory signal in the form of raw waveforms. TFEB replaces normal convolutions with a **Temporal Feature Extraction Block** made up of depthwise and pointwise convolutions that split filtering into spatial and channel parts. The overall design uses fewer redundant operations leading to a

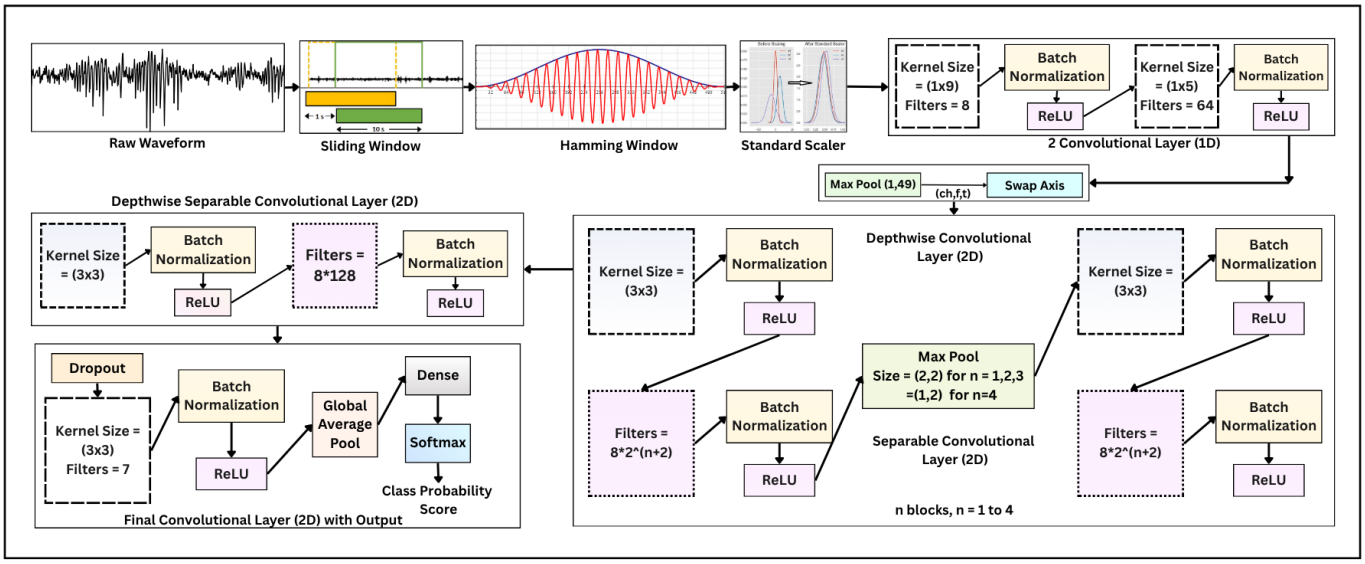


Fig. 2: Proposed DPNet Model Architecture with input, output and preprocessing

smaller model and faster inference, thus making it suitable for TinyML deployment.

### B. Model Complexity Analysis

The proposed architecture boosts the neural network's capability to a great extent as far as the parameters and the memory it needs are concerned. Standard convolution, however, has been substituted with depthwise-pointwise convolutions, which takes away the number of multiplications done in a layer by around 7 times.

TABLE I: Comparison of Model Complexity for Bathroom Sound Classification.

| Model | Trainable Parameters | Model Size (MB) |
|---|---|---|
| **ACLNet** | 148,352 | 1.30 |
| **ACDNet** | 4,711,303 | 36.9 |
| **DPNet (Proposed)** | 547,537 | 4.5 |

DPNet, as it can be seen from Table I, comes to a point of almost an **8× parameter reduction** when compared to ACDNet, and still, accuracy is not affected at all. The fact that memory is used less and inference is quicker, makes the application of such a network in low-power devices like ESP32-S3 Nano quite feasible.

### C. Training Details

The proposed bathroom sound dataset was the basis for the training of DPNet, which lasted 50 epochs in total with a batch size of 64. The specifications defined the use of a sliding window of 1.51125 s (hop length 0.51125 s) along with a Hamming window and feature standardization. The accuracy score of the model on the test set was 99.21%, and it also maintained this score after the quantization in TensorFlow Lite, thus validating its robustness for deployment on TinyML devices.

## VI. Experimental Results

### A. Performance Comparison with State-of-the-Art Models

DPNet was able to get an almost similar output to ACDNet, however it was much smaller and faster. Although ACDNet got the maximum offline accuracy, its size (36.9 MB) was a disadvantage for embedded deployment. While ACLNet occupied only 1.3 MB of space, the accuracy was also relatively low. Thus, DPNet was able to find an optimal combination of accuracy, model size and latency that made it suitable for TinyML-based real-time classification of sounds as it was the best among all three parameters.

### B. TFLite Quantization Model Assessment

The model sizes showed a dramatic reduction after converting to TensorFlow Lite. DPNet kept performing at a high level with its quantized versions—2.15 MB (*float32*, 99.21%), 1.09 MB (*float16*, 98.63%), and 0.68 MB (*int8*, 89.54%). On the other hand, ACDNet's *float32* model still was the biggest (18.5 MB) and ACLNet's *float32* TFLite model, though very small (584 KB), had only little real-world generalization.

### C. ESP32-S3 Nano Real-Time Deployment

An implementation of a dual-core processing scheme was done on the ESP32-S3 Nano: Core 0 took care of real-time PDM audio capture, while Core 1 did preprocessing and inference with the quantized TFLite model. The use of FreeRTOS queues combined with an OLED interface allowed the setup to operate in real-time within the 8 MB PSRAM limit of the board.

Despite the fact that the model size of ACLNet is drastically smaller (584 KB, *float32* TFLite), its real-time accuracy is quite low (22.64%) when compared to DPNet (73.44%) and ACDNet (42.78%). This points out the fact that although ACLNet is efficient in terms of power and size, the design

TABLE II: Performance Comparison of Models (Unquantized) on Original and Augmented Datasets

| Dataset | Models | Accuracy | Model Size (FP32) | Inference Time |
|---|---|---|---|---|
| Original (55 min/class) | ACLNet | 95.46% | 1.3 MB | 58.34 ms |
| | ACDNet | 99.39% | 36.9 MB | 63.75 ms |
| | Proposed **DPNet** | 97.52% | **4.5 MB** | **51.05 ms** |
| Original + Augmented (165 min/class) | ACLNet | 96.82% | 1.3 MB | 61.12 ms |
| | ACDNet | 99.64% | 36.9 MB | 66.67 ms |
| | Proposed **DPNet** | 99.21% | **4.5 MB** | **63.02 ms** |

TABLE III: Real-Time Performance of Quantized Models (which are deployable) on ESP32-S3 Nano

| Model | Quantization | Time (s) | Accuracy (%) |
|---|---|---|---|
| Proposed **DPNet** | Float32 | 11.18 | **73.44** |
| | Int16 | 8.14 | 60.58 |
| | Int8 | 6.92 | 48.32 |
| **ACDNet** | Int8 | 15.12 | 42.78 |
| **ACLNet** | Float32 | 12.84 | 22.64 |

is not so capable in dealing with complex and noisy bathroom settings. On the other hand, DPNet has managed to strike a balance between being compact and at the same time being accurate, thus, achieving the highest real-time performance not only among those compared models but also over the ESP32-S3 Nano platform.

## VII. Conclusion and Future Work

The present paper describes the DPNet model, which is a tiny but powerful deep learning model that is able to classify sounds in bathrooms in real-time even when running on microcontrollers with limited resources. Through the use of the combination of depthwise convolution and pointwise convolution in the Temporal Feature Extraction Block (TFEB), the size of the DPNet model is cut down drastically from 36.9 MB to 4.5 MB (FP32), and then further to 2.1 MB after quantization, with the accuracy being quite similar to that of the larger model. The float32 DPNet model was tested on the ESP32-S3 Nano and achieved a real-time accuracy of 73.44%, thus proving that it is a suitable solution for low-power acoustic sensing that preserves privacy and can be used in smart environments for the elderly.

In the course of time, we aim to upgrade our quantization methods so that we can eventually conquer memory usage without accuracy being affected, thus allowing to totally upload the int8 version within the 8 MB PSRAM of the ESP32-S3 Nano. Also, the application of adaptive noise filtering and online learning could make the system less sensitive to variations in the acoustic environment. The application of this framework to event detection across multiple rooms and the ability to learn from different domains would make it even more suitable for use in smart home and healthcare monitoring systems.

## References


[1] Salomons, E.L., Havinga, P.J. and Van Leeuwen, H., 2016. Inferring human activity recognition with ambient sound on wireless sensor nodes. Sensors, 16(10), p.1586.

[2] Chen, J., Kam, A.H., Zhang, J., Liu, N. and Shue, L., 2005, May. Bathroom activity monitoring based on sound. In International Conference on Pervasive Computing (pp. 47-61). Berlin, Heidelberg: Springer Berlin Heidelberg.

[3] Elliott, D., Otero, C.E., Wyatt, S. and Martino, E., 2021. Tiny transformers for environmental sound classification at the edge. arXiv preprint arXiv:2103.12157.

[4] Srivastava, S., Roy, D., Cartwright, M., Bello, J.P. and Arora, A., 2021, June. Specialized embedding approximation for edge intelligence: A case study in urban sound classification. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 8378-8382). IEEE.

[5] Lhoest, L., Lamrini, M., Vandendriessche, J., Wouters, N., da Silva, B., Chkouri, M.Y. and Touhafi, A., 2021. Mosaic: A classical machine learning multi-classifier based approach against deep learning classifiers for embedded sound classification. Applied Sciences, 11(18), p.8394.

[6] da Silva, B., W. Happi, A., Braeken, A. and Touhafi, A., 2019. Evaluation of classical machine learning techniques towards urban sound recognition on embedded systems. Applied Sciences, 9(18), p.3885.

[7] Maayah, M., Abunada, A., Al-Janahi, K., Ahmed, M.E. and Qadir, J., 2023. LimitAccess: on-device TinyML based robust speech recognition and age classification. Discover Artificial Intelligence, 3(1), p.8.

[8] Paranayapa, T., Ranasinghe, P., Ranmal, D., Meedeniya, D. and Perera, C., 2024. A comparative study of preprocessing and model compression techniques in deep learning for forest sound classification. Sensors, 24(4), p.1149.

[9] Mohaimenuzzaman, M., Bergmeir, C., West, I. and Meyer, B., 2023. Environmental Sound Classification on the Edge: A Pipeline for Deep Acoustic Networks on Extremely Resource-Constrained Devices. Pattern Recognition, 133, p.109025.

[10] Kadir, A.D.I.A., Al-Haiqi, A. and Din, N.M., 2021, December. A dataset and TinyML model for coarse age classification based on voice commands. In 2021 IEEE 15th Malaysia International Conference on Communication (MICC) (pp. 75-80). IEEE.

[11] Barovic, A. and Moin, A., 2025. Tinyml for speech recognition. arXiv preprint arXiv:2504.16213.

[12] Huang, Z., Tousnakhoff, A., Kozyr, P., Rehausen, R., Bießmann, F., Lachlan, R., Adjih, C. and Baccelli, E., 2024, September. TinyChirp: bird song recognition using TinyML models on low-power wireless acoustic sensors. In 2024 IEEE 5th international symposium on the internet of sounds (IS2) (pp. 1-10). IEEE.

[13] Fang, K., Xu, Z., Li, Y. and Pan, J., 2021, November. A fall detection using sound technology based on TinyML. In 2021 11th International Conference on Information Technology in Medicine and Education (ITME) (pp. 222-225). IEEE.

[14] Abadade, Y., Benamar, N., Bagaa, M. and Chaoui, H., 2024. Empowering healthcare: TinyML for precise lung disease classification. Future Internet, 16(11), p.391.

[15] Krishna, A., Shankaranarayanan, H., Oleti, H.P., Chauhan, A., Van Schaik, A., Mehendale, M. and Thakur, C.S., 2023, November. TinyML Acoustic Classification using RAMAN Accelerator and Neuromorphic Cochlea. In 2023 IEEE Asia Pacific Conference On Postgraduate Research In Microelectronics And Electronics (PRIMEAsia) (pp. 44-45). IEEE.

[16] Laput, G., Ahuja, K., Goel, M. and Harrison, C., 2018, October. Ubicoustics: Plug-and-play acoustic activity recognition. In Proceedings of the 31st annual ACM symposium on user interface software and technology (pp. 213-224).

[17] Tokozume, Y., Ushiku, Y. and Harada, T., 2017. Learning from between-class examples for deep sound recognition. arXiv preprint arXiv:1711.10282.

[18] Huang, J.J. and Leanos, J.J.A., 2018. Aclnet: efficient end-to-end audio classification cnn. arXiv preprint arXiv:1811.06669.