Title	NeuProNet: neural profiling networks for sound classification
Authors	Tran, Khanh-Tung;Vu, Xuan-Son;Nguyen, Khuong;Nguyen, Hoang D.
Publication date	2024-01-16
Original Citation	Tran, K. T., Vu, X. S., Nguyen, K. and Nguyen, H. D. (2024) 'NeuProNet: neural profiling networks for sound classification', Neural Computing and Applications, 36, pp. 5873-5887. https://doi.org/10.1007/s00521-023-09361-8
Type of publication	Article (peer-reviewed)
Link to publisher's version	10.1007/s00521-023-09361-8
Rights	© 2024, the Authors. Open Access. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. - https://creativecommons.org/licenses/by/4.0/
Download date	2026-08-09 06:33:42
Item downloaded from	https://hdl.handle.net/10468/15968

University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Neural Computing and Applications (2024) 36:5873–5887
https://doi.org/10.1007/s00521-023-09361-8

ORIGINAL ARTICLE

NeuProNet: neural profiling networks for sound classification

Khanh-Tung Tran[1] · Xuan-Son Vu[2] · Khuong Nguyen[1] · Hoang D. Nguyen[3]

Received: 2 March 2023 / Accepted: 7 December 2023 / Published online: 16 January 2024

Abstract
Real-world sound signals exhibit various aspects of grouping and profiling behaviors, such as being recorded from identical sources, having similar environmental settings, or encountering related background noises. In this work, we propose novel neural profiling networks (NeuProNet) capable of learning and extracting high-level unique profile representations from sounds. An end-to-end framework is developed so that any backbone architectures can be plugged in and trained, achieving better performance in any downstream sound classification tasks. We introduce an in-batch profile grouping mechanism based on profile awareness and attention pooling to produce reliable and robust features with contrastive learning. Furthermore, extensive experiments are conducted on multiple benchmark datasets and tasks to show that neural computing models under the guidance of our framework gain significant performance gaps across all evaluation tasks. Particularly, the integration of NeuProNet surpasses recent state-of-the-art (SoTA) approaches on UrbanSound8K and VocalSound datasets with statistically significant improvements in benchmarking metrics, up to 5.92% in accuracy compared to the previous SoTA method and up to 20.19% compared to baselines. Our work provides a strong foundation for utilizing neural profiling for machine learning tasks.

Keywords Neural profiling network · Audio classification · Deep learning · Signal processing

1 Introduction

With the advent of neural computing, sound classification systems have seen increasing use in the real world with a wide range of applications [1, 2], including in multimodal systems [3]. A practical application of these systems is to recognize health-related sounds like coughing and sneezing that can supply insights into the general well-being of each individual [4]. Such techniques can also detect significant lung sound anomalies such as wheezing and crackling, which previously could only be done by highly-skilled clinicians [5, 6]. Moreover, deploying machine-based sound surveillance systems is crucial as manufacturing and production lines become more autonomous towards predictive machine maintenance, boosting production quality and yield, and reducing downtime and costs [7]. Another imperative use case is automated environmental sound monitoring, such as ocean habitats [8] and urban noise sources. These analyses can provide a more detailed understanding of high pollution noise levels that affect our well-being [9].

Recent methods using neural networks have allowed efficient, cost-effective, and timely classification of various sound categories [10, 11]. Nevertheless, there are many drawbacks concerning the current state of the research direction. Whereas previous works have shown the advantages of applying machine learning techniques and deep neural networks to the sound domain, several insightful aspects of sound are yet to be investigated. First, sound signals can carry noisy and biased features of each original source and recording environment. These factors

✉ Hoang D. Nguyen
hn@cs.ucc.ie

Khanh-Tung Tran
tungtk9@fpt.com

Xuan-Son Vu
sonvx@cs.umu.se

Khuong Nguyen
khuongnd6@fpt.com

[1] AI Center, FPT Software Company Limited, Hanoi, Vietnam

[2] Department of Computing Science, Umeå University, Umeå, Sweden

[3] School of Computer Science and Information Technology, University College Cork, Cork, Ireland

can affect performances since models learn mixed characteristics linked to source information under diverse categories rather than downstream features [12–14]. Second, many existing methods consider each sound sample separately, even though they may come from identical sources, highly similar settings, or with the same recording backgrounds. Failing to utilize these signals in algorithms is more likely to lead to models missing complement features shared by the same provider and also being confused by irrelevant markers.

Prior works attempt to alleviate this by building various versions of the same systems for each specific device or patient [13, 15–17]. However, these methods have several limitations that may affect their applicability and effectiveness. First, they often rely on learning from each individual's data only, without information and weight sharing across different groups or individuals, requiring multiple training stages [13, 15, 16] and a satisfactory amount of data for each group [17]. Second, the learning capability and data quality of each group vary, which may lead to bias and fairness problems for some specific sets of groups. A third drawback is that existing methods are susceptible to reliability issues when a new user enters the system, as they may lack sufficient data or prior knowledge to provide accurate and relevant decisions [16]. To address existing limitations, we focus on a unified and end-to-end neural network solution that enjoys the benefit of personalization through neural profiling. As depicted in Fig. 1, we propose neural networks that learn from multiple audio samples that share a set of common attributes, instead of learning from each sample individually. These attributes can be based on the source, condition, location, or environment of the sound recordings. For example, it is feasible to group sounds from similar machines or humans, sounds from various conditions or backgrounds, or sounds from different places or situations.

Our main contributions are listed as follows:

- We propose novel neural profiling networks named NeuProNet capable of learning high-level personalized features through contrastive learning and in-batch profile grouping mechanisms. NeuProNet are designed to learn and extract reliable and robust profile embeddings from recordings in linked groups based on profile awareness and attention pooling.
- We develop an end-to-end framework to handle and demonstrate the effectiveness of NeuProNet when plugged with various backbone architectures on downstream audio-based tasks, including environmental sound classification and human vocal sound detection.
- We carry out extensive experiments on multiple benchmark datasets and tasks. We found that models under the guidance of neural profiling networks achieve significant performance gaps across all tasks. Particularly, our experimental results surpass recent state-of-the-art (SoTA) approaches with statistically significant improvements on UrbanSound8K, and VocalSound dataset with a substantial gap in benchmarking metrics, up to 20.19% in accuracy on UrbanSound8K.
- Source codes of our framework and reproducible baselines are made publicly available at https://github.

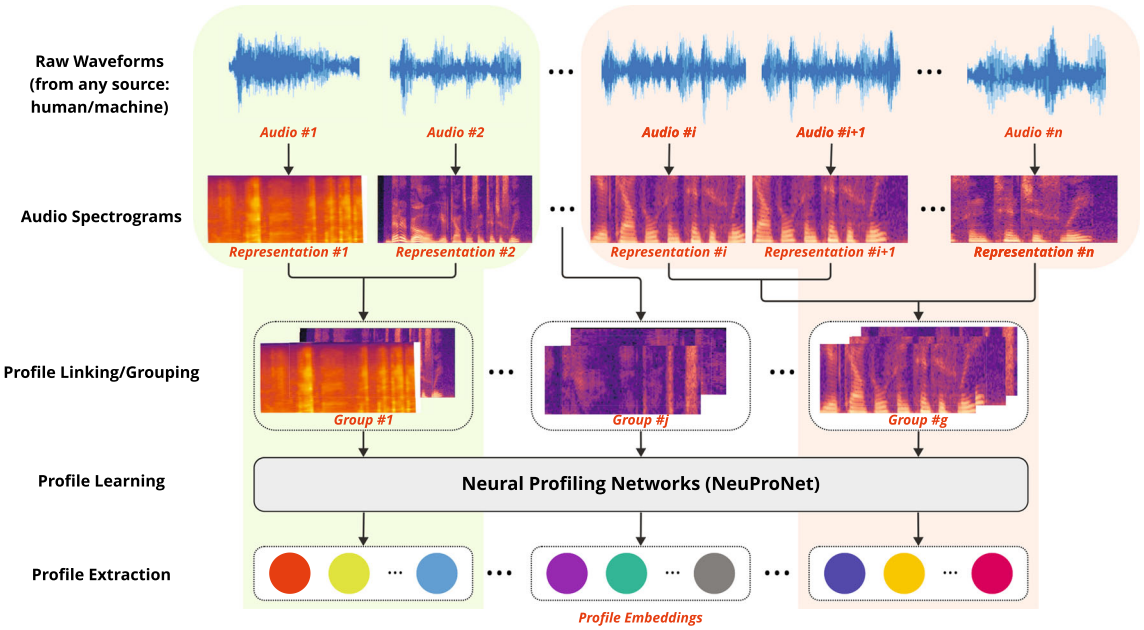

Fig. 1 We group each audio sample according to their grouping attribute, or their original source in this figure. Then, NeuProNet is employed to learn from multiple samples at once and extract the profile embedding that represents each group's unique and distinctive characteristics

com/ReML-AI/NeuProNet for future research and benchmarking purposes.

2 Background

In this section, we survey existing methods related to our works, summarize their strengths and weaknesses, and emphasize the novelty of our approach.

2.1 Sound & audio classification

In the general domain of audio pattern recognition, traditional systems comprised classical, statistical or machine learning models such as support vector machine [18]. Those systems would have time-frequency representations, e.g., log mel-spectrogram or mel-frequency cepstral coefficients, as input. This representation of data is also dominantly used for a large number of neural network techniques following the traditional machine learning approaches. Convolutional neural network-based systems are the most widely used [19–21] and outperformed recurrent neural network-based systems [22]. Recently, there has been a new direction of applying transformers with high complexities to this research domain. These transformer-based models enjoy flexibility in various types of input features, either directly as raw waveforms or the two-dimensional time-frequency representations [23, 24].

Since the amount of data for these types of datasets is typically small for deep learning, transfer learning and pre-training have also been applied [25, 26]. These techniques allow models to scale up to tens of millions of parameters. However, this also leads to massive computations, multiple training stages, and difficulty in representation learning for systems with real-world usages. Differing from them, our proposed mechanism that simultaneously learns personalized features for each origin-based group of samples with the typical downstream feature learning process alleviates the need for transfer learning. The method enjoys comparative performances and can be trained more efficiently and in an end-to-end manner.

The two sub-fields of environmental sound classification and vocal sound detection have seen a similar trend, as discussed above. However, most of the previous works only focus on either type of task. To our best knowledge, this is one of the pioneering works that bridge the gap between the two sub-domains.

2.2 From personalization to profiling

Personalization is successfully applied in numerous areas, such as in recommender and search systems [27] or medical applications [28]. In sound processing research, most personalized profiling works focus on speech personalization, including speech emotion recognition [29], speech enhancement [30], or humor detection [16]. A typical personalized machine learning method leverages additional cues, such as metadata related to each user. One other primary direction utilizes additional samples of the same user as the user embedding. For instance, Triantafyllopoulos et al. [29] proposed an enrolment-based method that learns on other voice samples of the present speaker to modify the main encoder's output for emotion recognition. In [31], the authors leveraged the speaker's noisy data that is commonly disregarded as pseudo-sources for a speech enhancement model. This work follows the second approach to fulfill privacy compliance without utilizing any additional metadata. Moreover, our profiling framework is applied to sound sources of both humans and surrounding habitats.

To the best of our knowledge, in sound pattern recognition, the work from Dang et al. [32] is the most closely connected study. Using samples from each user over time as a time-series and feeding them into an LSTM network, they examined the possibility of longitudinal audio samples over time for COVID-19 progression and recovery trend prediction. Regardless, they only worked on a private dataset with a binary task of COVID-19 detection from coughs. In this work, we develop a neural profiling network that can perform efficiently on a wide range of sound-related tasks.

3 Methodology

In this section, we present our proposed framework for neural profiling. The ultimate goal of our approach is to push the boundary of any neural network backbone on the downstream task by grounding each prediction to its respective profile embedding. As shown in Fig. 1, the framework consists of these steps:

- Profile linking/grouping: from each grouping attribute, such as recording device, speaker identifier, or recording environment, we gather audio samples for each source. Differing from prior works which deploy multiple grouping attributes, including sensitive ones, e.g., gender and age, to train models, we extract each source's profile embedding directly from sound samples, thereby profiling without sensitive information.
- Profile learning: deep neural networks are designed to learn and extract feature-rich profile embeddings from multiple samples in linked groups with contrastive and classification objectives.

- Profile extraction: our framework allows an end-to-end integration into any existing sound classification works for training and predicting with profile features in conjunction with audio features. Our approach improves neural computing systems even when the amount of available data is small or moderate.

This is enabled through the most essential component of the framework, NeuProNet, our neural network architecture developed to learn high-level profile representation through a candidate set of multiple audio samples. The overall architecture is depicted in Fig. 2.

3.1 Problem definition

Formally, define $\mathcal{D} = (X_i, \mathrm{g}_i, y_i)_{i=1}^{N_\mathcal{D}}$ as a dataset of $N_\mathcal{D}$ samples that are drawn i.i.d. from a data distribution. Each sample consists of an audio recording, denoted by $X_i \in \mathbb{R}^D$ for its raw waveform representation, or $X_i \in \mathbb{R}^{T \times F}$ for its spectrum form (e.g., spectrogram, which we utilize in this work). Additionally, each sample includes a group indicator $\mathrm{g}_i \in [\mathrm{G}]$, where G indicates the number of sources of data, and a class label $y_i \in \mathbb{R}$.

Typically, a learning algorithm adapts a generic classifier $h_\theta : \mathbb{R}^{T \times F} \rightarrow \mathbb{R}$ that maps from input variable X_i to class label $h_\theta(X_i)$. After the training phase, h_θ is able to perform a classification task with an empirical loss (or risk) on $\mathcal{D}$ as:

$$L(h_\theta, \mathcal{D}) = \frac{1}{N_\mathcal{D}} \sum_{i}^{i=N_\mathcal{D}} l(h_\theta(X_i), y_i) \tag{1}$$

where $N_\mathcal{D}$ denotes the total number of samples in the dataset $\mathcal{D}$.

Each individual who contributes to the final dataset expects the algorithm to tailor to them with respect to their group of samples and in turn earn a gain in performance. A tailored classifier $h_\theta^{\mathrm{g}} : \mathbb{R}^{T \times F} \times [\mathrm{G}] \rightarrow \mathbb{R}$ that uses group attributes in the best case will have:

$$L(h_\theta^{\mathrm{g}}, \mathcal{D}_\mathrm{g}) = \frac{1}{N_\mathrm{g}} \sum_{i:\mathrm{g}_i=\mathrm{g}} l(h_\theta^{\mathrm{g}}(X_i, \mathrm{g}_i), y_i) < L(h_\theta, \mathcal{D}_\mathrm{g}), \quad \forall \mathrm{g} \in [\mathrm{G}] \tag{2}$$

where N_g denotes the number of samples belonging to group g.

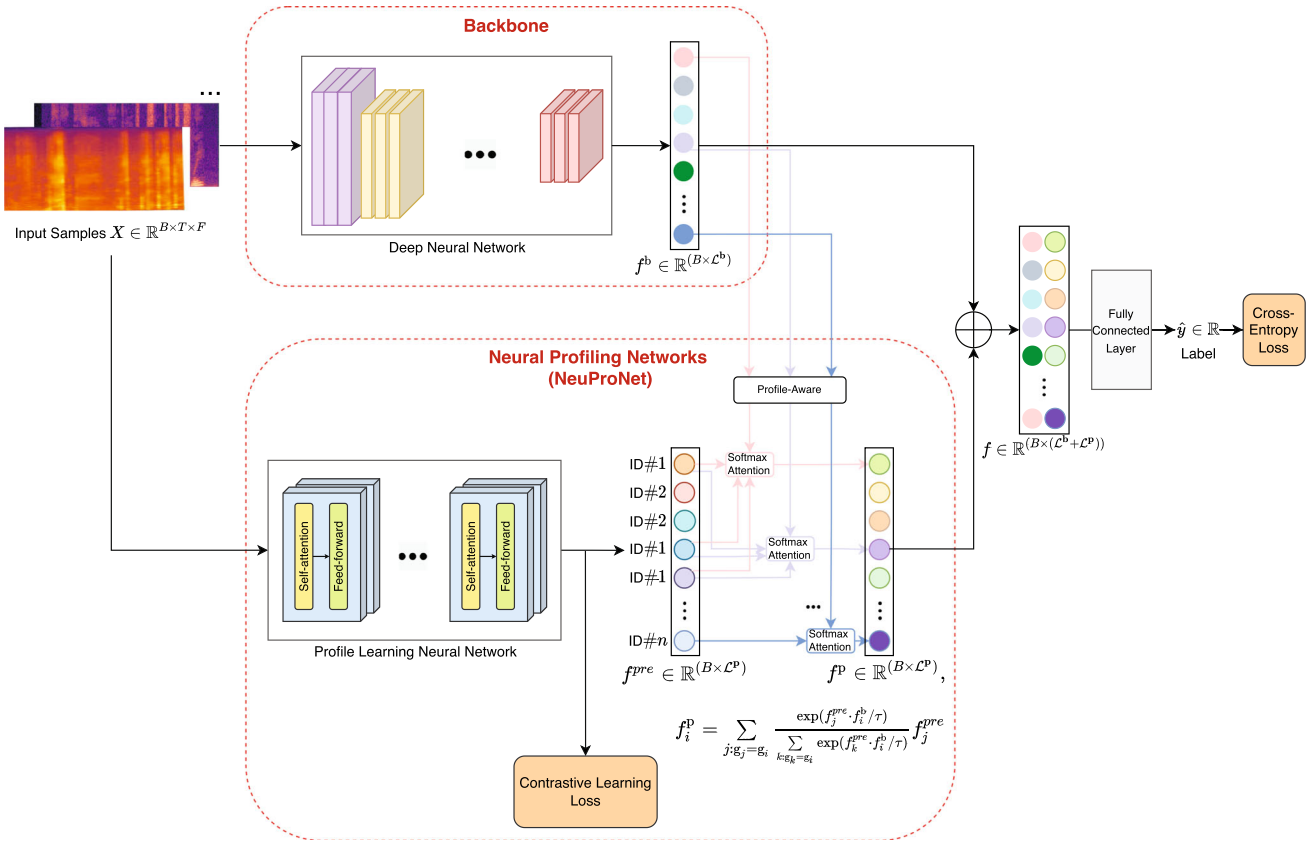

Fig. 2 Our neural profile learning framework. Batched-input spectrograms of shape $B \times T \times F$ are input to both branches of the framework simultaneously. The first branch leverages a common approach in sound classification, typically using a CNN-based or transformer-based model. The second branch is our method of profile learning, NeuProNet. The final output of the two branches is concatenated together (depicted by the $\oplus$ operator) before being passed to a classification layer

Leading to:

$$L(h_\theta^{\mathrm{g}}, \mathcal{D}) < L(h_\theta, \mathcal{D}) \tag{3}$$

This means that by incorporating additional group features into the learning process, the classifier achieves better overall performance on the dataset $\mathcal{D}$.

In conventional practice, a set of model parameters for each group is trained particularly on that group's dataset. This forces the same model architecture to be saved in several iterations, and each model is thus unable to make use of a larger dataset. Another drawback is that sensitive metadata (such as sex or health condition) must be used in order to understand each group's features successfully. As discussed in the following sections, we proposed a neural profile learning framework that enjoys the advantage of profiling in one single unified model based only on sound data without sensitive metadata.

3.2 Neural profiling network

In our proposed framework for neural profiling, the goal is to learn unique features for each group of samples. This is achieved through the neural profile learning mechanism by analyzing each group's provided samples, which are called candidate samples with respect to the current sample for the downstream task. As mentioned in the previous section, an effective analysis of the candidate set can help a tailored classifier gain better performance.

Our proposed approach leverages an additional neural network that is dedicated to learning from groups of samples. It differs from the backbone branch where each input sample is computed separately. By introducing another branch to the system concurrently with the backbone branch, the backbone network with the existing structure and optimal settings can remain intact and be distilled from the neural profile learning process. Additionally, our approach maximizes embeddings learned and shared by the backbone branch with its dedicated neural architecture. This flexibility allows any backbone network to be plugged into the system.

Moreover, we also analyze deeply and realize that the candidate set may come from any class of the target task, and if handled poorly, it can contribute additional noise to the model. An ineffective profile learning strategy would confuse the model and negatively impact performance. Previous works tend to require another external dataset as the set of candidate samples to avoid confusion. To address this issue, we propose extracting these tailored features directly from input samples rather than collecting them from another corpus, which is a common practice in previous works. This approach reduces the need and cost to collect additional data and allows for more efficient learning of new representations from the same data. Intuitively, by combining embeddings learned by two branches in the framework, we can potentially achieve better performance by leveraging the strengths of each representation and compensating for their weaknesses, as each representation captured certain aspects of the data.

Feature vector extracted by the backbone network in the traditional manner:

$$f_i^{\mathrm{b}} = h_\theta^{\mathrm{b}}(X_i) \tag{4}$$

Combine with the neural profiling branch:

$$f^{\mathrm{p}} = \frac{1}{N_{\mathrm{g_i}}} \sum_{j:\mathrm{g}_j=\mathrm{g}_i} h_\theta^{\mathrm{p}}(X_j) \tag{5}$$

Thereby, the final prediction for sample i is described below:

$$\hat{y_i} = \mathcal{W}(f_i^{\mathrm{b}} \oplus f_i^{\mathrm{p}}) \tag{6}$$

where $\mathcal{W}$ denotes a linear classifier head and $\oplus$ denotes the concatenate operation. This allows the two branches to learn and update concurrently. Each branch is proposed to learn from different input data: the backbone network is designed to learn signals from an audio sequence, where NeuProNet is proposed to learn from a group of multiple audio sequences. Therefore, deploying both branches concurrently instead of stacking enables the backbone network to retain its original architecture while introducing a separate network for neural profiling. We can utilize the benefits of both branches for the downstream task, as the representations learned from the backbone branch and the profile learning branch are different and can complement each other. Whereas there are studies on more sophisticated methods for fusing representations between multiple branches [33], we employ a simple concatenation operator, as we aim for our proposed NeuProNet to be a new SoTA baseline and as easy as possible to be applied to any pre-existing networks. Nevertheless, our experiments and analyses still show good signals from both branches and we expect performance can be further improved if more advanced fusion techniques are applied. In our analysis of the worst-case scenario, the model will still be able to perform adequately by focusing on the backbone based on our fully connected linear mechanism.

Our framework enables both neural network branches to extract separate representations and to be trained concurrently from the classification tasks. To help information sharing between the two branches, we derive the use of a softmax attention pooling mechanism, where softmax scores of dot products between the sample's embedding from the backbone branch to its corresponding candidate set's profiling features (profile-aware). This allows the system to filter out noisy features and highlight valuable information from the neural profiling branch, improving

not only the profiling learning task but also the backbone branch's learning process:

$$f_i^{\text{pre}} = h_\theta^{\text{p}}(X_i),$$
$$f_i^{\text{p}} = \sum_{j:\text{g}_j=\text{g}_i} \frac{\exp(f_j^{\text{pre}} \cdot f_i^{\text{b}}/\tau)}{\sum_{k:\text{g}_k=\text{g}_i} \exp(f_k^{\text{pre}} \cdot f_i^{\text{b}}/\tau)} f_j^{\text{pre}} \tag{7}$$

where exp is the exponential function, and τ is the temperature hyperparameter.

Moreover, while we design our framework to be able to plug into any backbone architectures, NeuProNet leverages a Transformer-based network for the neural profiling branch. This is due to the highly efficient attention mechanism of the transformer model and its freedom from inductive biases. Additionally, for learning meaningful profiles, we employ a supervised contrastive learning loss [34], where the labels used to group positive and negative samples are obtained through group attributes, e.g., patient ID, recording environment, etc., where negative samples are not chosen from samples of the same group so they can be pushed closer together.

$$L_{\text{CL}} = -\log \frac{\exp(\text{sim}(f_i^{\text{pre}}, f_{j,\text{g}_j=\text{g}_i}^{\text{pre}}))}{\sum_{k:\text{g}_k \sim [\text{G}]\setminus \text{g}_i} \exp(\text{sim}(f_i^{\text{pre}}, f_k^{\text{pre}}))} \tag{8}$$

where sim is a similarity function such as the cosine similarity.

The model is trained in an end-to-end manner through back-propagation via a cross-entropy loss function. Both branches can be updated through gradient descent.

$$L_{\text{CE}} = \sum_i y_i \log(\sigma(\hat{y_i})) + (1 - y_i) \log(1 - \sigma(\hat{y_i})) \tag{9}$$

where $\sigma(\cdot)$ is known as the activation function.

The final loss function for the NeuProNet branch is:

$$L_{NeuProNet} = \lambda L_{\text{CE}} + (1 - \lambda) L_{\text{CL}} \tag{10}$$

where λ is treated as a tuned hyperparameter.

Note that in the final system, however, one would retrieve candidate samples from the entire dataset or another corpus, leading to slow training and inference time. To alleviate this, we propose to group samples solely on samples available in the current batch of input data. The candidate set, on average, may be reduced in size. However, our ablation study shows that with this mechanism, our proposed approach achieves the same level of performance while enjoying sufficient speed in both stages.

4 Experiments

This section details the experimental procedures, datasets, evaluation metrics, and implementation designs utilized to evaluate the proposed network's performance and benchmark it against baselines and SoTA approaches.

4.1 Datasets and evaluation metrics

We validate the proposed framework on two different publicly available sound classification datasets, namely UrbanSound8K and VocalSound. A detailed description of these datasets is presented below.

4.1.1 UrbanSound8K

The UrbanSound8K dataset [35] includes 8732 mono and stereo samples divided into ten categories: air conditioner, car horn, children playing, dog barking, drilling, engine idling, gunfire, jackhammer, siren, and street music. The number of samples in each class is varied, making this an imbalanced dataset. Moreover, the average recording durations for each class are not distributed evenly throughout the courses. Each track may be up to 4 s long and has a native sampling rate ranging from 16 to 48 kHz. As a result, the collection provides a more accurate reconstruction of daily ambient noises, including both machine- and human-based sounds. Other urban datasets are less suited since they are more predictable and less "naturalistic".

We construct each group, or each candidate set, as samples from the same original recorded audio before they are split, as indicated in the metadata contained in the dataset. While the original Freesound recording contains signals from different classes at different time stamps, each split audio provided in the dataset consists of a signal for the class and various background noises related to each recording environment, i.e. inside or outside a house. These noisy signals can be common between each group's samples.

Table 1 Hyperparameters used in the experimental setup

Hyperparameter	Value on UrbanSound8K	Value on VocalSound
Sampling rate	22.05 kHZ	22.05 kHZ
Learning rate	1e−4	1e−4
Optimizer	AdamW	AdamW
Number of training epochs	100 (3500 with augmentation)	50
Batch size	512	128

The authors of the dataset separated it into ten folds, which we employed in this study to conduct our evaluation. As discussed in previous works [35, 36], to avoid data leakage, it is of crucial importance to follow the official split. Therefore, we only consider prior works that perform experiments on the official split of UrbanSound8K. We do not perform comparisons with customized splits such as in [37] because those splits are not reproducible. Moreover, while our framework is not limited to transfer learning, pre-training, or multimodal training paradigms, we focus on fair comparisons with other end-to-end methods on audio data.

As some classes in the dataset have fewer samples than others, we report both accuracy and AUC, along with other metrics, to account for the issue of data imbalance.

4.1.2 VocalSound

VocalSound [12] is a publicly accessible dataset containing 21,024 crowdsourced audio recordings of laughter, sighs, coughs, throat clearing, sneezes, and sniffs from 3365 unique persons. Meta information about each speaker, namely anonymous speaker label, speaker age, gender, native language, nationality, and health status, are also included. The dataset focuses on research on removing bias on various speaker groups, as the authors acknowledged an EfficientNet [38]-based model has biases among certain groups.

We leverage group samples provided by each person as a candidate set. In general, each speaker provided around six non-speech sounds, each belonging to a separate class. This setting also fits with applications such as patient health monitoring, where anomaly signals like cough and sneeze can be detected over a duration span and can be used as indicators of the general well-being of individuals.

We follow the evaluation strategy of the original paper, where the dataset is split into three sets, training, validation, and testing set. Additionally, we repeat experiments on each model three times and report the mean and standard deviation of evaluation statistics, with accuracy as the main benchmarking metric.

4.2 Benchmarking methods

As we are interested in the efficacy of the neural profiling framework, we compare various backbone networks before and after NeuProNet is plugged in. These implementations are heavily inspired by SoTA audio classification models, given their popularity and ease of reproducibility.

- Resnet: we deploy a vanilla Resnet18 [39] model, a variant of the CNN which uses residual blocks. While originally proposed for tasks in the computer vision domain, recent research has focused on effectively applying these types of neural networks with residual connection to the domain of audio signal processing through the spectrogram representation of sound.
- EfficientNet: EfficientNet-B0 [38] is a recently proposed convolutional neural network that is also based on an efficient residual convolution block. EfficientNet has the ability to scale the network on all dimensions (width, depth, and input resolution). This model has been shown to obtain high performances in sound classification tasks.
- Transformer: there has been a surge of research in applying transformer architecture for audio data. In this work, we use a model that stacks multiple layers of transformer-encoder blocks together.

Moreover, we also compare results with previous SoTA methods that performed benchmarking on the same (fair) data splits. Our framework not only achieves significant gaps in performances across all tasks and datasets compared to backbones without NeuProNet, but it also allows

Table 2 Comparison of results against various baseline methods on UrbanSound8K for settings without augmentation

Method	Accuracy	Precision	Recall	F1 score	AUC ROC	#Para[$\times 10^6$]
MC-DCNN (2019) [40]	75.1	–	–	–	–	–
LMCNet (2019) [41]—reproduced by [36]	74.04	–	–	–	–	15.9
AUCO Resnet (2022) [42]	77.83	78.51	77.83	77.09	96.77	22.1
EfficientNet	63.56 ± 3.73	66.97 ± 4.42	65.79 ± 3.95	64.76 ± 4.16	91.91 ± 1.86	5.5
NeuProNet+EfficientNet	***83.75*** ± ***3.74***	***86.04*** ± ***2.71***	***84.98*** ± ***2.99***	***84.77*** ± ***3.07***	***97.25*** ± ***1.06***	9.5
Resnet	74.55 ± 4.48	77.26 ± 5.62	74.85 ± 3.83	73.98 ± 4.74	94.99 ± 1.59	7.3
NeuProNet+Resnet	***78.46*** ± ***5.61***	***81.03*** ± ***5.84***	***78.80*** ± ***4.67***	***78.46*** ± ***5.49***	***95.67*** ± ***1.47***	11.3
Transformer	75.69 ± 2.77	76.70 ± 2.70	77.29 ± 2.79	75.91 ± 2.67	95.24 ± 1.29	4.0
NeuProNet+Transformer	***81.53*** ± ***2.49***	***83.51*** ± ***2.05***	***82.16*** ± ***2.39***	***81.98*** ± ***2.23***	***97.31*** ± ***0.92***	8.0

We only consider prior works that perform experiments on the official split of UrbanSound8K. Best values are highlighted in bold

Table 3 Comparison of results against various baseline methods on UrbanSound8K for settings with augmentation

Method	Accuracy	Precision	Recall	F1 score	AUC ROC	#Para[$\times 10^6$]
ESResnet-Att (2021) [36] ♢	82.76	–	–	–	–	32.6
ERANN-1–4 (2022) [19] ♢	83.5	–	–	–	–	24.1
EfficientNet ★	67.49 ± 3.61	74.12 ± 2.87	66.91 ± 4.55	67.09 ± 3.71	92.83 ± 1.84	5.5
NeuProNet+EfficientNet ★	***81.56*** ± ***5.31***	***85.99*** ± ***3.80***	***80.72*** ± ***5.11***	***81.11*** ± ***5.37***	***96.31*** ± ***1.75***	9.5
Resnet ★	72.46 ± 4.03	76.18 ± 4.00	72.09 ± 4.06	71.94 ± 4.32	94.67 ± 1.91	7.3
NeuProNet+Resnet ★	***81.90*** ± ***5.18***	***84.79*** ± ***4.26***	***81.34*** ± ***5.18***	***81.74*** ± ***4.91***	***96.49*** ± ***1.75***	11.3
Transformer ★	74.86 ± 4.72	76.36 ± 4.70	75.00 ± 4.83	74.70 ± 5.05	95.20 ± 1.96	4.0
NeuProNet+Transformer ★	***83.34*** ± ***4.24***	***84.93*** ± ***4.48***	***82.78*** ± ***4.44***	***82.92*** ± ***4.63***	***96.95*** ± ***1.59***	8.0

♢ denotes that the results were obtained by finetuning models pre-trained on other datasets

★ denotes our solutions with augmentation during training phase

We only consider prior works that perform experiments on the official split of UrbanSound8K. Best values are highlighted in bold

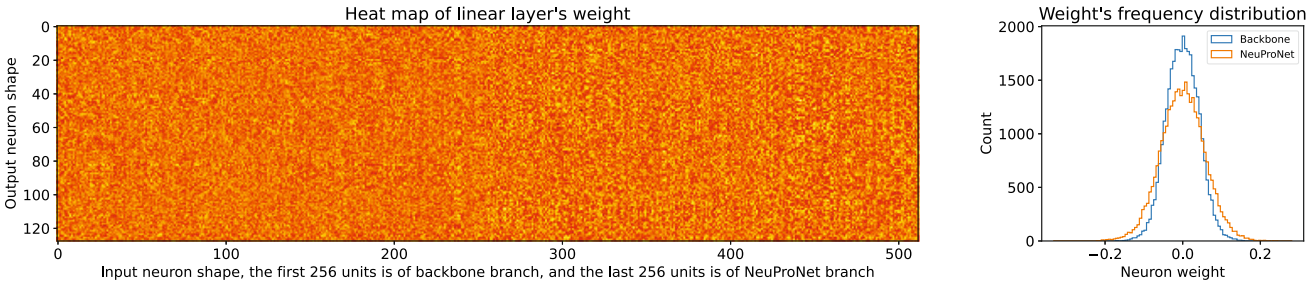

Fig. 3 Visualization of weights of the linear layer immediately after feature's concatenation between the two branches in our framework for experiments on UrbanSound8K

these naive backbone architectures to obtain comparable or even better performances than previous works.

In order to assess the proposed approach and compare it with the existing works, accuracy is our primary evaluation metric. Other standard evaluation metrics such as precision, recall, F1, and AUC are also reported for future benchmarking purposes.

4.3 Training and implementation details

The proposed framework was developed in PyTorch, and experiments were carried out on an NVIDIA A100 GPU. Detailed information about our hyperparameter settings is provided in Table 1. These values are chosen based on previous studies and were kept consistent across the training baselines, whether with or without NeuProNet. Regarding the λ value in 10, it is set to 0.9999 for experiments on UrbanSound8K or to 0.99 on VocalSound, based on the scale of the contrastive learning loss on each dataset. The other values are set empirically through grid search. As mentioned in Sect. 4.1, we adhere to the official data splits and benchmarking strategy specified by the authors of each dataset. To ensure fair comparisons, we also adhered to the data augmentation techniques and weight imbalancing methods used in previous works. All model trainings were completed in under 2 h, and there was no additional overhead in training or inference time when NeuProNet was integrated. This is because NeuProNet operates in parallel with the backbone branch.

5 Results and discussion

5.1 Results on UrbanSound8K

In our study, we perform experiments both with or without augmentation to ensure reliability and reproducibility; because some concerned data augmentation can greatly reduce the reproducibility of the processes. The results are illustrated in Tables 2 and 3.

Popular backbones (e.g., EfficientNet, Resnet, and Transformer) are thoroughly investigated from the public domain of existing deep learning approaches, without any customization for the audio domains. Without the use of NeuProNet, these backbones individually achieve lower performances than the current SoTA without augmentation. Therefore, our NeuProNet approach demonstrates great

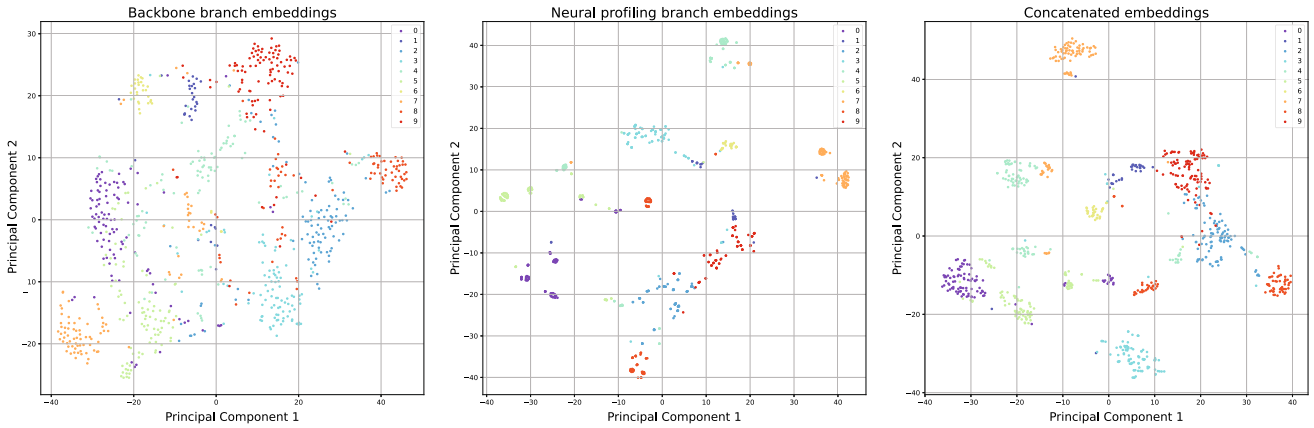

Fig. 4 t-SNE visualizations of learned embeddings on UrbanSound8K dataset. Legends on the corner denote the class labels. The embeddings of the backbone branch are illustrated in the left panel, and the embeddings of the neural profiling branch are illustrated in the middle panel. Finally, concatenated embeddings of the two branches are visualized in the right panel

Table 4 Comparison of other metrics besides the benchmarking metric (accuracy) on VocalSound for settings without augmentation

Method	Accuracy	Precision	Recall	F1 score	AUC ROC	#Para[$\times 10^6$]
EfficientNet	91.02 ± 0.34	91.05 ± 0.34	91.02 ± 0.33	91.02 ± 0.33	98.57 ± 0.07	5.3
NeuProNet+EfficientNet	***91.81*** ± ***0.30***	***91.82*** ± ***0.33***	***91.80*** ± ***0.31***	***91.80*** ± ***0.32***	***98.83*** ± ***0.01***	9.3
Resnet	91.23 ± 0.42	91.29 ± 0.42	91.23 ± 0.42	91.22 ± 0.42	98.82 ± 0.11	7.3
NeuProNet+Resnet	***92.34*** ± ***0.15***	***92.40*** ± ***0.11***	***92.34*** ± ***0.14***	***92.33*** ± ***0.14***	***99.01*** ± ***0.05***	11.3
Transformer	89.12 ± 0.85	89.25 ± 0.74	89.12 ± 0.86	89.14 ± 0.82	***98.47*** ± ***0.11***	4.0
NeuProNet+Transformer	***90.42*** ± ***0.63***	***90.35*** ± ***0.61***	***90.35*** ± ***0.61***	***90.33*** ± ***0.61***	96.99 ± 0.36	8.0

Best values are highlighted in bold

Table 5 Comparison of results on VocalSound for settings with augmentation

Method	Accuracy	Precision	Recall	F1 score	AUC ROC	#Para[$\times 10^6$]
EfficientNet (2022) [12]	90.5 ± 0.20	–	–	–	–	5.3
EfficientNet ★	92.99 ± 0.05	93.01 ± 0.05	92.99 ± 0.05	92.98 ± 0.04	99.27 ± 0.01	5.3
NeuProNet+EfficientNet ★	***93.91*** ± ***0.25***	***93.94*** ± ***0.26***	***93.91*** ± ***0.25***	***93.91*** ± ***0.25***	***99.50*** ± ***0.02***	9.3
Resnet ★	92.86 ± 0.20	92.86 ± 0.20	92.86 ± 0.20	92.84 ± 0.20	99.22 ± 0.04	7.3
NeuProNet+Resnet ★	***93.98*** ± ***0.41***	***93.94*** ± ***0.48***	***93.92*** ± ***0.50***	***93.93*** ± ***0.49***	***99.43*** ± ***0.07***	11.3
Transformer ★	86.06 ± 0.27	86.25 ± 0.22	86.06 ± 0.27	86.00 ± 0.28	97.93 ± 0.01	4.0
NeuProNet+Transformer ★	***88.42*** ± ***0.47***	***88.55*** ± ***0.48***	***88.47*** ± ***0.51***	***88.46*** ± ***0.53***	***98.44*** ± ***0.09***	8.0

- ★ denotes our solutions with augmentation during training phase

Best values are highlighted in bold

improvements with large gaps, regardless of the backbone's characteristics and architecture. The reason for this reliability and robustness is due to our neural profiling mechanism that helps learn novel representations of data in addition to information learned by the backbone branch. Notably, we observe a massive boost in the accuracy of the EfficientNet model, up to 20.19% compared against backbones without the neural profiling network. This was proved to be statistically significant in our neural profiling method (two-tailed t-test; $p < 0.01$). Furthermore, when compared with SoTA solutions, our NeuProNet approach allows simple backbone architectures to outperform all other approaches on the dataset without any specific modifications and achieve a gap of 5.92% between the EfficientNet method with neural profiling network and previous SoTA—AUCO Resnet [42]. The accuracy of

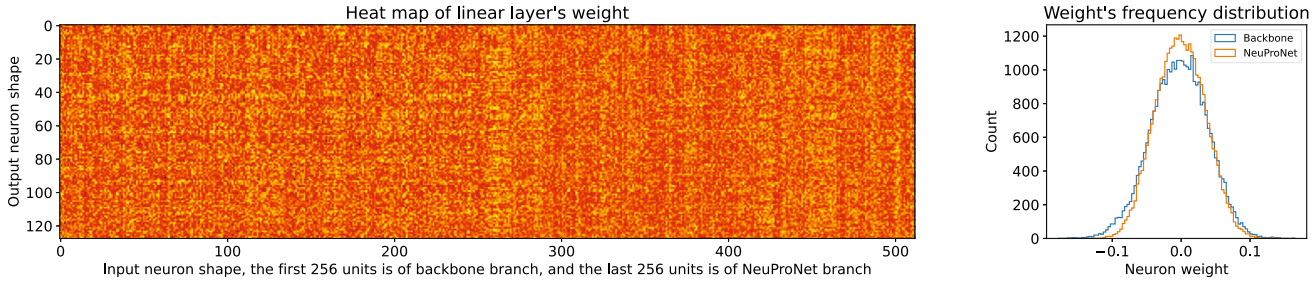

Fig. 5 Visualization of weights of the linear layer immediately after feature's concatenation between the two branches in our framework for experiments on VocalSound

83.75% of NeuProNet+EfficientNet is also comparable with the SoTA method with augmentation. Our framework accomplishes these novel results while maintaining lower complexity in terms of the number of trainable parameters.

With data augmentation, the training process becomes even more costly. Following common approaches, we deploy SpecAugment [43] as our augmentation method in training time. SpecAugment is used for frequency and time masking on spectrograms of training audio clips. It can be seen that augmentation creates perplexities in recognizing the right patterns of classes, leading to degraded performances in 2 out of 3 backbone models (Resnet and Transformer). However, with the addition of the neural profiling network, we achieve even higher performances for both NeuProNet+Resnet and NeuProNet+Transformer with respect to the setting without augmentation. This indicates the robustness against random added noises and confusions of our proposed mechanism, where traditional backbones failed. Consistent with results without augmentation, all three backbones achieve a substantial gap when NeuProNet is plugged in, up to 14.07%. Moreover, the highest accuracy of 83.34% obtained on the Transformer model with NeuProNet is comparable with other SoTA results, with a much less powerful backbone architecture and more straightforward augmentation strategy, allowing for better reproducibility.

We visualize the weights of the first linear layer that takes the concatenation feature vector of the backbone branch and NeuProNet branch to analyze the contribution of each branch to the downstream task. Figure 3 shows the heatmap of the layer on the left, and the frequency distribution of the weight values, split by weights that are directly related to each branch. Looking at the heatmap, we can see that weights associated with NeuProNet are reasonably more activated, and the frequency distribution indicates that features learned by NeuProNet have higher effects, with more weights of the linear layer associated with NeuProNet have values near the tail of the distribution than that of the backbone branch. This is proof that NeuProNet has a greater impact on the framework's final decision than the backbone.

All the results indicate the efficacy of the neural profiling network in surpassing previous methods through learning new information and signals from the candidate set for environmental sound classification on UrbanSound8K. We also compute other metrics among the benchmarking metric (accuracy), namely precision, recall, F1 score, and AUC ROC. The results are reported in Tables 2 and 3 for settings without or with augmentation, respectively. Generally, our proposed neural profiling network helps obtain better performances across all metrics with large gaps. This proves the generalizability of our method without overfitting into the benchmarking metric.

To gain a better understanding of the learned embeddings of each branch and the concatenated embeddings, we used t-distributed stochastic neighbor embedding (t-SNE) to obtain a two-dimensional visualization, as shown in Fig. 4. It can be seen clearly that the embeddings of the neural profiling branch provide different signals that are clearer and stronger than those of the backbone branch. While embeddings of the backbone architecture have shown signs of clustering into each class, each cluster is sparse and tends to overlap. However, the concatenation embeddings of the backbone and neural profiling branches demonstrated a greater distance between each cluster and each one of them is denser than previously. This allows for more reliable performance in classification.

5.2 Results on VocalSound

As we analyzed in the prior section, this dataset may introduce a level of confusion to typical personalization strategies, where the candidate set is drawn from different classes than the input sample. We are interested in how our framework overcomes these hardnesses of the dataset. Results are presented in Tables 4 and 5. To our best knowledge, this is one of the first works to perform benchmarking on this dataset. Therefore, in this section, we will focus on a comparison with the baseline method accompanied with the dataset, which is based on EfficientNet-B0 architecture for their efficiency, as indicated in Sect. 4.2. We deploy a re-implementation of their method

Table 6 Comparison of accuracy on each sub-group of users that contributed samples to the test set on VocalSound for settings without augmentation

Method	Age group			Variance between age groups↓	Gender group		Variance between gender groups↓
	18 – 25 ↑	26 – 48 ↑	49 – 80 ↑		Male↑	Female↑	
EfficientNet	91.59	90.71	91.28	***0.45***	89.76	92.33	1.82
NeuProNet+EfficientNet	***92.56***	***91.55***	***92.43***	0.55	***90.92***	***93.00***	***1.47***
Resnet	91.97	90.83	91.76	0.61	89.98	92.49	1.77
NeuProNet+Resnet	***92.57***	***92.16***	***92.43***	***0.21***	***91.23***	***93.54***	***1.63***
Transformer	87.93	89.05	89.37	0.76	86.75	89.20	1.73
NeuProNet+Transformer	***89.75***	***90.19***	***89.97***	***0.22***	***88.71***	***90.38***	***1.18***

Table 7 Comparison of accuracy on each sub-group of users that contributed samples to the test set on VocalSound for settings with augmentation

Method	Age Group			Variance between age groups↓	Gender group		Variance between gender groups↓
	18 – 25 ↑	26 – 48 ↑	49 – 80 ↑		Male↑	Female↑	
EfficientNet (2022) [12]	91.5	90.1	90.9	0.70	89.2	91.9	1.91
EfficientNet ★	93.91	92.70	93.20	0.61	91.94	93.92	1.40
NeuProNet+EfficientNet ★	***94.51***	***93.82***	***93.68***	***0.44***	***93.02***	***94.82***	***1.27***
Resnet ★	93.50	92.36	94.25	0.95	91.71	94.02	1.63
NeuProNet+Resnet ★	***94.51***	***93.61***	***94.83***	***0.63***	***93.05***	***94.82***	***1.25***
Transformer ★	84.68	87.24	83.43	1.94	84.71	88.52	2.69
NeuProNet+Transformer ★	***88.57***	***88.44***	***88.25***	***0.16***	***87.34***	***89.61***	***1.60***

★ denotes our solutions with augmentation during training phase

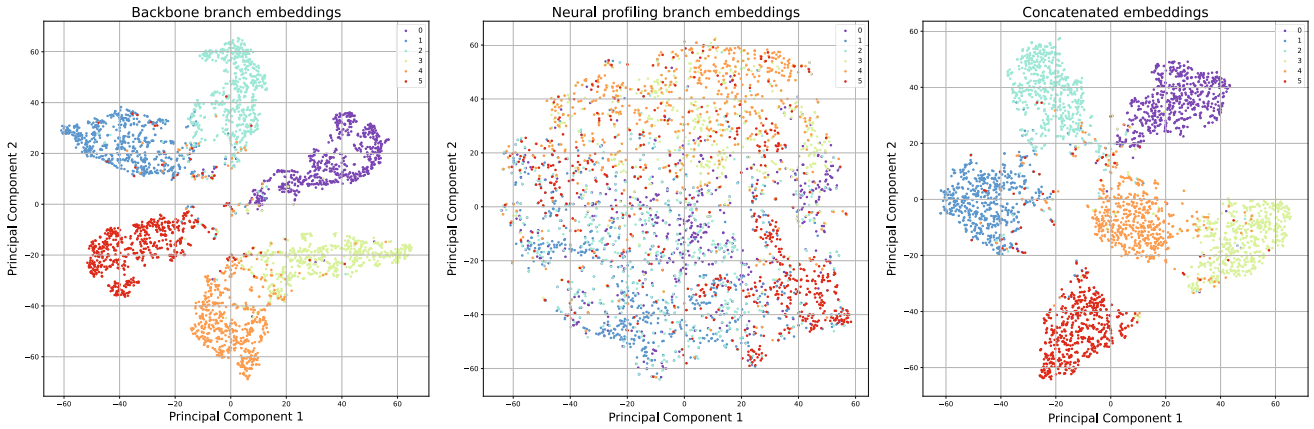

Fig. 6 t-SNE visualizations of learned embeddings on VocalSound dataset. Legends on the corner denote the class labels. The embeddings of the backbone branch are illustrated in the left panel, and the embeddings of the neural profiling branch are illustrated in the middle panel. Finally, concatenated embeddings of the two branches are visualized on the right panel

for fairer comparisons. We follow the original training setup of the dataset's paper, using SpecAugment as the augmentation strategy. Moreover, we also conduct training experiments without augmentation for comparisons on more scenarios.

With respect to training without augmentation, as indicated in Table 4, performances of backbone methods

Table 8 Results of ablation study where we set out to understand contributions of each component of our proposed NeuProNet framework

Components				UrbanSound8K	VocalSound
softmax attention pooling	contrastive loss	profile grouping mechanism	NeuProNet		
✗	✗	✗	✗	75.69 ± 2.77	89.12 ± 0.85
✗	✗	✗	✔	75.47 ± 2.70	90.02 ± 0.37
✗	✗	✔	✔	81.15 ± 2.83	90.12 ± 0.18
✗	✔	✔	✔	81.44 ± 2.44	90.16 ± 0.53
✔	✔	✔	✔	***81.53*** ± ***2.49***	***90.42*** ± ***0.63***

without neural profiling mechanism are varied compared to that of training with augmentation. While the accuracies of Effnet and Resnet models decreased, the accuracy of the backbone Transformer model increased. However, NeuProNet still maintains a positive gap when they are plugged in, up to 1.30% in the case of the Transformer-based model. The improvement earned by plugging in NeuProNet when training with augmentation is comparable to training without augmentation. This is proof that our proposed approach is more robust and not affected by augmentation, as seen in backbone networks. These phenomena are also seen in Table 4. A potential explanation is that NeuProNet is capable of learning valuable and robust embeddings from the candidate set to support decisions in the downstream task.

On setting with augmentation, we achieve a new SoTA result with a high accuracy score of 93.98% or 93.91% when neural profiling networks are combined with Resnet or Effnet backbone, respectively. Our highest accuracy score beats the previous baseline of the dataset by 3.48%, indicating the effectiveness of the proposed profiling strategy. Table 5 also clearly shows that our proposed solutions with NeuProNet outperform their backbone versions consistently across all metrics and across all backbone model architectures. The most significant difference is seen when transformers are chosen as the backbone network, where our proposed neural profiling network helps improve performance by 2.36% in accuracy score, from 86.06 to 88.42%. This result is statistically significant (two-tailed t-test, $p < 0.05$).

One observation is that the improvement of NeuProNet is more significant on UrbanSound8K than VocalSound. It suggests that the hardness of VocalSound, where each candidate set is drawn from different classes than the input sample, affects the features learned by the neural profiling network to support downstream tasks. In other words, it is harder to extract useful information from the candidate set related to the current sample's label when they are drawn from different classes. Moreover, the performances of baseline and backbones on each dataset indicate that VocalSound might be easier to solve than UrbanSound8K, and embeddings learned by the backbone models are able to incorporate more information related to the task. Therefore, the embeddings learned by the neural profiling branch are able to help in fewer cases. This is also verified by looking at the visualization of weights of the linear layer with input as the concatenation of embeddings learned by the two branches, illustrated in Fig. 5. We can see that the distribution of weights dedicated for each branch is more similar to each other, and the heatmap shown on the left is more evenly distributed than on UrbanSound8K, indicating that both branches contribute more equally to the final prediction.

Furthermore, we investigate possible bias in our profiling networks, as there has been an ongoing concern of bias in the evaluation of classification results with respect to each speaker group [12]. Table 7 shows that their method was biased towards younger audiences and towards female over male speakers. Nevertheless, our proposed neural profile learning approach takes into account each speaker's individual profile based on the candidate set, allowing all speakers to benefit from the system regardless of age or gender, as shown in Tables 6 and 7. The implications of those experiments are twofold. First, between each pair of the same backbone network architecture, the system with neural profiling networks obtains better performance across all groups of speakers. Secondly, the variance between each set of groups is much lower for solutions with neural profiling networks, with the most significant gap in the Transformer model for the setting with augmentation, from 1.94 to 0.16 variance between age groups, and from 2.69 to 1.60 for gender groups. This indicates that NeuProNet helps to achieve more reliable and fair classification results.

The above insights evidently show that, in general, NeuProNet helps obtain better performance, even though confusion in the candidate set may affect the performance of the proposed network. Nevertheless, we show that our

proposed neural profiling network always improves performance regardless of datasets and models.

We also use t-SNE to visualize learned embeddings of each branch and concatenated embeddings on VocalSound dataset, as illustrated in Fig. 6. Similar to the results on UrbanSound8K, in the embeddings of the backbone architecture, each class's cluster still has overlapping areas. However, with the support of the neural profiling framework, the concatenation embeddings have highly separated and denser clusters.

5.3 Ablation study

In the ablation study, we carry out experiments to further analyze and understand the effectiveness of the proposed NeuProNet. Every component of the framework is plugged out one after another, and the results are illustrated in Table 8. Here, we show results with Transformer as the backbone branch for experiments on both UrbanSound8K and VocalSound, but the results are generalized across different backbone architectures. While it is clear that all modules contributed to the gain in performance of the proposed framework, the profile grouping mechanism, or the method to ensemble features from the candidate set, is the most valuable component. On the other hand, the setting without the grouping mechanism achieves equal performance to the backbone network. This means that in the case of unavailable group labels, the framework does not hurt the model's performance.

6 Conclusion

Overall Conclusion. We propose a novel method for sound classification, namely NeuProNet, a neural profiling network capable of learning latent profile representation shared between audio samples from identical sources. Differing from prior works, we explore the capability of neural networks in a new learning regime, with an in-batch profile grouping mechanism and contrastive loss on group attributes to learn high-level profile embedding. NeuProNet can be plugged in with any backbone architecture, e.g., EfficientNet, Resnet, and Transformer, as well as, trained in an end-to-end manner. Through extensive experiments and analyses, we show that NeuProNet learn and extract robust and reliable features, obtaining high performances consistently across different backbone networks and datasets. We evaluate our framework on multiple evaluation metrics, namely accuracy, precision, recall, F1 score, and ROC AUC. In particular, we achieve substantial improvement up to 5.92% in accuracy compared to the previous SoTA approach and up to 20.19% compared to the baseline. Moreover, by producing profile representation, NeuProNet allows all sound sources, e.g., speakers, to benefit from the system, regardless of their meta attributes, such as gender or age, resulting in reliable and fair classification results that surpass bias problems in previous approaches. Future works may be interested in further exploring and analyzing latent profile characteristics or more advanced fusion methods between backbone features and profiles to be less dependent on backbone performances. Another interesting research direction is to apply training paradigms such as pre-training and semi-supervised learning to learn profile features from large and unlabeled data sets.

Practical Implications. The proposed NeuProNet and neural profiling framework have the potential to significantly improve the performance and efficiency of various sound-related applications, benefiting a wide range of industries and domains, such as acoustic system monitoring and surveillance, environmental sound classification, and sound anomaly detection. By extracting high-level profile representation for each machine based on their id (specific-unit) or machine class (specific type), NeuProNet helps achieve more accurate performance in device anomaly detection, reducing false alarms and maintenance costs. Furthermore, our framework can also help in sound surveillance systems. As more samples become available, the longer the system runs, profile representation extracted by NeuProNet becomes more accurate and robust to noise, leading to even higher performance in downstream tasks. Overall, the proposed framework has the potential to enhance the reliability and effectiveness of sound-related applications across various domains.

Broader Impact. In this work, we develop a neural profile learning framework that allows for any backbone network to be combined with our proposed neural profiling network and gain significant improvement in various sound-related tasks. Moreover, we also show that these improvements are consistent and scalable regardless of backbone architecture or dataset. In other words, any pre-existing systems can be plugged into our framework and enjoy robust and high performances without consequences.

Funding Open Access funding provided by the IReL Consortium.

Data availability The UrbanSound8K dataset can be found at: https://doi.org/10.1145/2647868.2655045 [35]. The VocalSound dataset can be found at: http://doi.org/10.1109/ICASSP43922.2022.9746828 [12].

Declarations

Conflict of interest There is no conflict of interest.

References

1. Herremans D, Chuan CH (2019) The emergence of deep learning: new opportunities for music and audio technologies. Neural Comput Appl 32(4):913–914
2. Coelho G, Matos LM, Pereira PJ, Ferreira A, Pilastri A, Cortez P (2022) Deep autoencoders for acoustic anomaly detection: experiments with working machine and in-vehicle audio. Neural Comput Appl 34(22):19485–19499
3. Sharma A, Sharma K, Kumar A (2022) Real-time emotional health detection using fine-tuned transfer networks with multimodal fusion. Neural Comput Appl 35(31):22935–22948
4. Imran A, Posokhova I, Qureshi HN, Masood U, Riaz MS, Ali K et al (2020) AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. Inform Med Unlocked 20:100378
5. Earis J, Cheetham B (2000) Current methods used for computerized respiratory sound analysis. Eur Respir Rev 01(10):586–590
6. Rocha BM, Filos D, Mendes L, Vogiatzis I, Perantoni E, Kaimakamis E et al (2018) A respiratory sound database for the development of automated classification. In: Maglaveras N, Chouvarda I, de Carvalho P (eds) Precision medicine powered by pHealth and connected health. Springer Singapore, Singapore, pp 33–37
7. Bukhsh Z (2022) Contrastive sensor transformer for predictive maintenance of industrial assets. In: ICASSP 2022—2022 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 3558–3562
8. Williams B, Lamont TAC, Chapuis L, Harding HR, May EB, Prasetya ME et al (2022) Enhancing automated analysis of marine soundscapes using ecoacoustic indices and machine learning. Ecol Ind 140:108986
9. Raimbault M, Dubois D (2005) Urban soundscapes: experiences and knowledge. Cities 22(5):339–350
10. Panda R, Malheiro RM, Paiva RP (2020) Audio features for music emotion recognition: a survey. IEEE Trans Affect Comput 14:68–88
11. Chandrakala S, Jayalakshmi SL (2019) Environmental audio scene and sound event recognition for autonomous surveillance: a survey and comparative studies. ACM Comput Surv 52(3):1–34
12. Gong Y, Yu J, Glass J (2022) Vocalsound: a dataset for improving human vocal sounds recognition. In: ICASSP 2022—2022 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 151–155
13. Gairola S, Tom F, Kwatra N, Jain M (2021) Respirenet: a deep neural network for accurately detecting abnormal lung sounds in limited data setting. In: 2021 43rd annual international conference of the IEEE engineering in medicine & biology society (EMBC). IEEE, pp 527–530
14. Han J, Xia T, Spathis D, Bondareva E, Brown C, Chauhan J et al (2022) Sounds of COVID-19: exploring realistic performance of audio-based digital testing. NPJ Digit Med 5(1):1–9
15. Acharya J, Basu A (2020) Deep neural network for respiratory sound classification in wearable devices enabled by patient specific model tuning. IEEE Trans Biomed Circuits Syst 14(3):535–544
16. Kathan A, Amiriparian S, Christ L, Triantafyllopoulos A, Müller N, König A, et al (2022) A personalised approach to audiovisual humour recognition and its individual-level fairness. In: Proceedings of the 3rd international on multimodal sentiment analysis workshop and challenge. MuSe' 22. Association for Computing Machinery, New York, NY, USA, pp 29–36
17. Kathan A, Harrer M, Küster L, Triantafyllopoulos A, He X, Milling M et al (2022) Personalised depression forecasting using mobile sensor data and ecological momentary assessment. Front Digit Health 4:964582. https://doi.org/10.3389/fdgth.2022.964582
18. Wei P, He F, Li L, Li J (2019) Research on sound classification based on SVM. Neural Comput Appl 32(6):1593–1607
19. Verbitskiy S, Berikov V, Vyshegorodtsev V (2022) ERANNs: efficient residual audio neural networks for audio pattern recognition. Pattern Recogn Lett 161:38–44
20. Pham L, Ngo D, Tran K, Hoang T, Schindler A, McLoughlin I (2022) An ensemble of deep learning frameworks for predicting respiratory anomalies. In: 2022 44th annual international conference of the IEEE engineering in medicine & biology society (EMBC), pp 4595–4598
21. Nguyen T, Pernkopf F (2022) Lung sound classification using cotuning and stochastic normalization. IEEE Trans Biomed Eng 69(9):2872–2882
22. Li J, Dai W, Metze F, Qu S, Das S (2017) A comparison of deep learning methods for environmental sound detection. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 126–130
23. Gong Y, Chung YA, Glass J (2021) AST: audio spectrogram transformer. In: Proceedings of Interspeech 2021, pp 571–575
24. Chen K, Du X, Zhu B, Ma Z, Berg-Kirkpatrick T, Dubnov S (2022) HTS-AT: a hierarchical token-semantic audio transformer for sound classification and detection. In: ICASSP 2022—2022 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 646–650
25. Gong Y, Chung YA, Glass J (2021) PSLA: improving audio tagging with pretraining, sampling, labeling, and aggregation. IEEE/ACM Trans Audio Speech Lang Process 29:3292–3306
26. Wang Z, Wang Z (2022) A domain transfer based data augmentation method for automated respiratory classification. In: ICASSP 2022—2022 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 9017–9021
27. Zhou Y, Dou Z, Zhu Y, Rong Wen J (2021) PSSL: self-supervised learning for personalized search with contrastive sampling. In: Proceedings of the 30th ACM international conference on information & knowledge management, pp 2749–2758
28. Weiss JC, Natarajan S, Peissig PL, McCarty CA, Page D (2012) Machine learning for personalized medicine: predicting primary myocardial infarction from electronic health records. AI Mag 33(4):33
29. Triantafyllopoulos A, Liu S, Schuller BW (2021) Deep speaker conditioning for speech emotion recognition. In: 2021 IEEE international conference on multimedia and expo (ICME), pp 1–6
30. Eskimez SE, Yoshioka T, Wang H, Wang X, Chen Z, Huang X (2022) Personalized speech enhancement: new models and comprehensive evaluation. In: 2022 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 356–360

31. Sivaraman A, Kim S, Kim M (2021) Personalized speech enhancement through self-supervised data augmentation and purification. In: Proceedings of the Interspeech 2021
32. Dang T, Han J, Xia T, Spathis D, Bondareva E, Brown C et al (2022) Exploring longitudinal cough, breath, and voice data for COVID-19 disease progression prediction via sequential deep learning: model development and validation (preprint). J Med Internet Res 02:24
33. Hazarika D, Zimmermann R, Poria S (2020) Misa: modality-invariant and-specific representations for multimodal sentiment analysis. In: Proceedings of the 28th ACM international conference on multimedia, pp 1122–1131
34. Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P et al (2020) Supervised contrastive learning. Adv Neural Inf Process Syst 33:18661–18673
35. Salamon J, Jacoby C, Bello JP (2014) A dataset and taxonomy for urban sound research. In: 22nd ACM international conference on multimedia (ACM-MM'14). Orlando, FL, USA, pp 1041–1044
36. Guzhov A, Raue F, Hees J, Dengel A (2021) ESResNet: environmental sound classification based on visual domain models. In: 2020 25th international conference on pattern recognition (ICPR). IEEE Computer Society, Los Alamitos, CA, USA, pp 4933–4940
37. Al-Hattab YA, Zaki HF, Shafie AA (2021) Rethinking environmental sound classification using convolutional neural networks: optimized parameter tuning of single feature extraction. Neural Comput Appl 33(21):14495–14506
38. Tan M, Le Q (2019) Efficientnet: rethinking model scaling for convolutional neural networks. In: International conference on machine learning. PMLR, pp 6105–6114
39. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
40. Chong D, Zou Y, Wang W (2019) Multi-channel convolutional neural networks with multi-level feature fusion for environmental sound classification. In: Kompatsiaris I, Huet B, Mezaris V, Gurrin C, Cheng WH, Vrochidis S (eds) MultiMedia modeling. Springer International Publishing, Cham, pp 157–168
41. Su Y, Zhang K, Wang J, Madani K (2019) Environment sound classification using a two-stream CNN based on decision-level fusion. Sensors 19(7):1733
42. Dentamaro V, Giglio P, Impedovo D, Moretti L, Pirlo G (2022) AUCO ResNet: an end-to-end network for Covid-19 pre-screening from cough and breath. Pattern Recogn 127:108656
43. Park DS, Chan W, Zhang Y, Chiu CC, Zoph B, Cubuk ED, et al (2019) SpecAugment: a simple data augmentation method for automatic speech recognition. Interspeech 2019. Sep