Evaluating Transfer Learning Strategies for Lung Sound Event Classification

Michiel Jacobs[1,2,3,4,5][0000−0002−7920−0659], Lode Vuegen[1,2,3][0000−0002−3418−4069], and Peter Karsmakers[1,2,3][0000−0001−8119−6823]

[1] KU Leuven, Dept. of Computer Science, Kleinhoefstraat 4, 2440 Geel, Belgium
[2] Flanders Make @ KU Leuven
[3] Leuven.AI - KU Leuven Institute for AI
[4] Hasselt University, Limburg Clinical Research Center, Mobile Health Unit, 3500 Hasselt, Belgium
[5] Ziekenhuis Oost-Limburg, Future Health, 3600 Genk, Belgium
{michiel.jacobs, peter.karsmakers}@kuleuven.be

Abstract. Computerised lung auscultation has the potential to offer automated respiratory disease follow-up in ambulatory settings. Lung sound recordings are typically analysed using Sound Event Classification (SEC) models. However, during inference, mismatches between the training and deployment data distributions can lead to significant performance degradation. Transfer Learning (TL) techniques offer a way to mitigate this problem.
In this study, we evaluate SEC performance on two in-house lung sound datasets using: (a) models trained on publicly available lung sound data, and (b) those models enhanced with domain+task TL, domain TL and semi-supervised domain+task TL methods. We conclude that, for our setup, domain TL results in good classification performance when only a domain shift is present. When a task shift exists between source and target data, partially labelled target data is required to obtain good task adaptation.

Keywords: Adventitious lung events · sound event classification · domain adaptation · transfer learning · DANN.

1 Introduction

In recent years, there has been a growing interest in detecting adventitious events from lung sounds, e.g. crackles and wheezes [5, 7, 8]. Crackles are explosive sounds and typically last between 5 and 15 ms. Wheezes are musical in nature with frequencies in the range of 100 to 5000 Hz, with a typical duration of more than 100 ms [1, 11].

Lung sound datasets used for training Sound Event Classification (SEC) models typically include recordings from multiple auscultation positions and a variety of stethoscope devices. These datasets also feature recordings from numerous participants, each with a unique physique that acts as an individual

acoustic filter [1], introducing further variability in the captured signals. Two well-known public datasets that will be used in this study are ICBHI [11] and HFLungV1 [4].

However, publicly available training datasets are generally limited in size and diversity, making it unlikely that they capture the full range of variability encountered in real-world scenarios. As a result, there is a significant risk that the domain (defined as the feature distribution of the data) of a target dataset will differ, at least slightly, from that of the source domain training data. This discrepancy is commonly referred to as domain shift.

To address such domain shifts between source and target datasets, Transfer Learning (TL) techniques can be employed. In their foundational work, Pan et al. [10] introduced two key concepts to distinguish TL strategies, i.e. domain and task transfer learning. As defined above, a domain consists of a feature space and a marginal probability distribution over that space, while a task comprises a label space and a predictive function that maps input features to output labels. Transfer learning may involve transferring knowledge across domains, tasks, or both.

In this work, both the task (auscultation vs. autogenic drainage therapy, and different pathologies) and the domain (different stethoscopes) may vary. We follow the definitions in [10], referring to domain TL as the setting where labelled source data and unlabelled target data are available, and domain+task TL as the case where both source and target datasets are labelled. Semi-supervised domain+task TL then refers to adding unsupervised learning to domain+task TL, using both source data and (partially) labelled target data.

Domain TL techniques have shown promise in improving lung SEC under domain shift. For instance, Kim et al. [8] proposed a stethoscope-guided supervised contrastive learning method. In this approach, each type of stethoscope is treated as a separate domain, and contrastive learning is used to map similar events to nearby regions in the latent space regardless of the recording device. When evaluated on the ICBHI dataset, their method achieved a 2.16% improvement in average recall, reaching 61.17% in total. Similarly, Hsu et al. [5] found that models trained on lung sounds recorded from the chest performed poorly when applied to tracheal recordings and vice versa. To address this, they applied TL and mixed-set training. Both techniques improved performance, with mixed-set training yielding the largest gains. In another study, Huang et al. [6] introduced the Contrastive Embedding-Based Domain Adaptation Neural Network (CEDANN) to distinguish between healthy children and those with pneumonia. On unseen patients, CEDANN increased sensitivity from 54.74% to 64.17% and specificity from 58.44% to 68.05%.

These studies demonstrate the potential of TL to improve lung sound classification performance under domain and task shift. In this study, we compared the use of domain+task TL, domain TL, and a semi-supervised domain+task TL algorithm across two different public source datasets (ICBHI and HFLungV1) and two in-house target datasets containing recordings from patients with Chronic Obstructive Pulmonary Disease (COPD) and Cystic Fibrosis (CF). Under the

assumption of correctly labelled target samples, domain+task TL -which requires annotation of the target data- can be viewed as an upper bound benchmark, and semi-supervised domain+task TL -which is less reliant on labelled target data- is therefore more practical in real-world applications where labelling is costly or infeasible.

The following sections are organised as follows: Section 2 outlines the TL methods that were used, as well as the evaluation metric chosen. Section 3 then defines the experimental setup. Next, Section 4 describes the results obtained. Section 5 discusses these results and aims to give explanations on these results. Finally, Section 6 summarises all work into a conclusion.

2 Methodology

2.1 Transfer Learning

In this section three types of Transfer Learning (TL) are compared, i.e. domain+task TL, domain TL, and semi-supervised domain+task TL.

In domain+task TL, the model is first trained on source data and then fine-tuned on target data. This process allows the model to adapt to potential domain shifts [10]. Moreover, the model can be specialised for a slightly modified task, such as a redefined event class.

For the domain TL approach, the Domain Adversarial Neural Network (DANN) framework proposed by Ganin et al. [3] was employed. In this setup, a domain discriminator was added as a secondary output head to the feature extractor network. Its objective is to identify the domain (here dataset) from which each input sample originated. During backpropagation, the discriminator's gradient is negated before being combined with the classifier's gradient, and the result is backpropagated to the feature extractor. The gradient negation encourages the feature extractor to learn domain-invariant representations, i.e. aligning the source and target feature distributions. Similar to [12], we normalised the negated gradient of the discriminator on the feature space, rescaled it with the norm of the gradient of the classifier and multiplied it with a constant strictly greater than 1. This result is backpropagated into the feature extractor, prioritising domain-invariant features over good classification performance. The advantage of this is that the hyperparameter λ from the original DANN is eliminated. This approach is summarised in Equation 1. As an unsupervised approach, DANN does not require labelled data from the target data. However, DANN cannot account for class redefinitions (task changes), as opposed to domain+task TL.

$$\nabla_{W_e} L\left(x, y_C, y_D\right) := \nabla_{W_e} L_C\left(x, y_C\right) - 1.5 \frac{\nabla_{W_e} L_D\left(x, y_D\right)}{\left\|\nabla_{\mathcal{F}_e} L_D\left(x, y_D\right)\right\|} \left\|\nabla_{\mathcal{F}_e} L_C\left(x, y_C\right)\right\| \tag{1}$$

where W_e the weights of the feature extractor, $\mathcal{F}_e$ the feature extractor's output, and L_C and L_D refer to the classifier loss and domain discriminator loss, respectively.

Semi-supervised domain+task TL combines elements of both the domain+task TL and unsupervised DANN approaches. When both a domain shift and a task shift are expected, unsupervised methods are insufficient. In such cases, a limited amount of labelled target data is required to adapt the model to the modified task. By combining the unsupervised DANN strategy with domain+task TL, the number of necessary labelled target samples could potentially be reduced.

Figure 1 summarises all TL methods considered.

2.2 Evaluation

Area Under the Receiver Operating Characteristic curve (AUROC) was chosen as the evaluation metric. Preliminary studies showed that decision thresholds vary between patients. Therefore, AUROC was chosen as the evaluation metric since it assesses the classifier's ranking performance and is not dependent on a decision threshold.

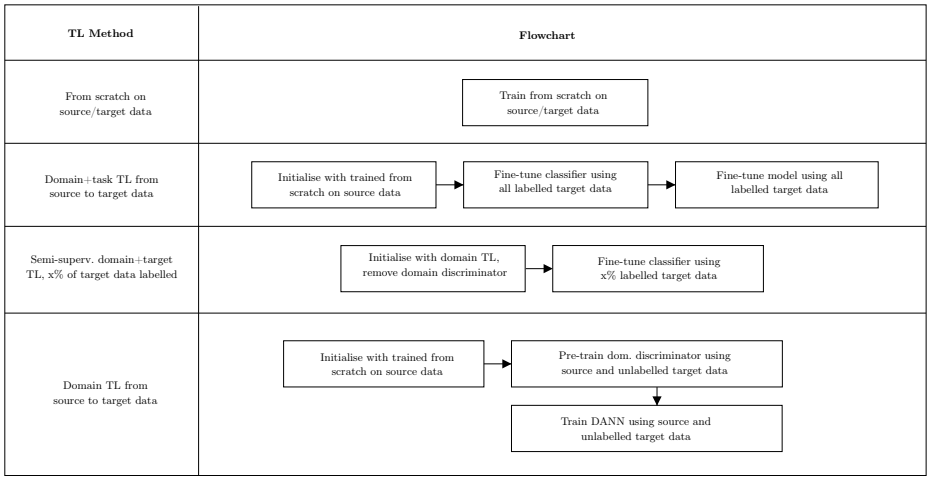

Fig. 1. Flowcharts visualising the various TL setups tested.

3 Experiments

3.1 Datasets

The ZOL in-house dataset was collected at Ziekenhuis Oost-Limburg (ZOL) hospital using a Littmann 3200 stethoscope and contains 346 1-minute auscultation recordings from 8 patients suffering from Chronic Obstructive Pulmonary Disease (COPD). It can be subdivided into two sets. The first set was composed by a trained annotator and was later reviewed by a physician. These data contain 5

patients (300 recordings) and were split into 5 training and validation folds with equal patient distributions across the folds. The second set was composed by majority vote over three trained annotators and contains 3 patients (46 recordings). This second set served as our independent test set. Due to the patient pathology only wheezing events are present.

The Compass in-house dataset was collected at the University Hospital Brussels (UZ Brussel). It contains data from 5 patients with either COPD or Cystic Fibrosis (CF) during autogenic drainage therapy. Trained annotators labelled the audio for crackles, rales, crepitations, and wheezes. Crackles, rales and crepitations were then grouped into one class "crackles" as these are discontinuous adventitious sounds, and to be consistent with the event classes in the other datasets. Leave-one-patient-out cross-validation was used to create 5 test folds. The remaining 4 patients formed the training and validation sets. Data were recorded using a ThinkLabs One stethoscope on the chest anterior auscultation positions.

The ICBHI public dataset [11] was first presented at the International Conference on Biomedical Health Informatics (ICBHI) in 2017. In total, 920 recordings (5.5 h) were recorded from 126 participants suffering from various respiratory diseases and across various auscultation positions. Four unique stethoscopes were used, each with their own sampling rate. We divided the official training set into 5 cross-validation folds by sampling the recordings in a stratified way such that the distribution of patients across folds was equal. The official test set was not used in this work.

The HFLungV1 public dataset [4] was collected by F. Hsu, S. Huang, C. Huang et al. In total, 9,765 audio files were recorded, each with a duration of 15 s. Data were recorded with either a Littmann 3200 stethoscope or with a Heroic Faith Type 1 device at various auscultation positions. Both devices have a sampling rate of 4000 Hz.

If a lung sound dataset includes tracheal recordings, these data are excluded due to their distinct channel characteristics, which differ significantly from those of chest and back recordings [1, 5]. Therefore, this study focusses exclusively on lung sounds recorded over the chest and back.

Table 1 summarises the stethoscopes that occur in each of the datasets.

Table 1. Stethoscopes used in each dataset.

↓ Dataset, Stethoscope →	AKG C417L	HF Type 1	Littmann 3200	Littmann C2SE	Meditron	ThinkLabs One
ICBHI	67%		1%	16%	16%	
HFLungV1		58%	42%			
Compass						100%
ZOL			100%			

3.2 Pre-Processing

In a first step, all audio were resampled to 4000 Hz, since most adventitious lung events have frequencies well below 2000 Hz [1]. As a second step, a twelfth-order Butterworth high-pass filter with cut-off frequency equal to 60 Hz was used

to limit the interference of heart beats. For each stethoscope and auscultation position, time-domain audio amplitudes were rescaled such that the maximum amplitude across recordings of the same stethoscope and auscultation position is equal to 1. Third, 1 s audio chunks were converted to spectrograms using Short-Time Fourier Transform (STFT) with 25 ms windows and 10 ms steps. This resulted in 65 frequency bins. The Hann window function is applied. The magnitude spectrogram was converted to log-power scale. When needed, zero padding was applied at the end of the audio recording.

3.3 Model & Training

The Convolutional Neural Network (CNN) architecture from previous work [7] was improved. The CNN consists of 4 convolutional blocks. Each of these blocks performs a 2D convolution operation, followed by `ELU` activation [2]. Next, 2×2 max pooling is performed. The first two blocks have 5×5 convolutional kernels, while the last two blocks have 3×3 kernels. The multi-label classifier contains two hidden dense layers with `ReLU` activation. The first hidden dense layer has 128 neurons, and the second hidden dense layer has 32 neurons. Weight decay was set to 0.01. The batch size was always equal to 256 samples. Dropout was applied on all layers with a drop rate of 50%. Multi-label output was applied for crackle and wheeze classification. Since all datasets suffer class imbalances, a stratified batch sampler was always used that yields the same class distribution inside each batch of data. The `AdamW` optimiser was used [9].

For domain TL and semi-supervised domain+task TL, a discriminator network is necessary to classify the domain an input sample originates from. This discriminator network consists of 2 hidden dense layers of 32 neurons each. Both layers have `ReLU` activation. Finally, the output layer consists of 1 neuron with sigmoid activation.

3.4 Transfer Learning

With domain+task TL, the model was first pre-trained on publicly available lung sound data. In this work, we opted for ICBHI and HFLungV1 lung sound datasets as the source datasets. Next, the model was fine-tuned on our in-house ZOL and COMPASS target datasets. During this fine-tuning, only the classifier was trained during the first 50 epochs. Afterwards, the entire model was fine-tuned for 200 epochs.

In our implementation of the unsupervised DANN network, the domain discriminator consisted of two fully connected hidden layers, each with 32 neurons and `ReLU` activation. The output layer contained a single neuron with sigmoid activation to perform binary domain classification (source vs. target). The training process began by pre-training the discriminator for 50 epochs to reliably distinguish between source and target domains. Afterwards, the feature extractor was updated jointly using both the classifier and negated discriminator gradients. This setup will be referred to as "domain TL".

We also tested a setup in which the best unsupervised DANN was further trained using varying portions of labelled target data. Here, only the classifier was trained; the weights of the feature extractor were kept constant. The domain discriminator was removed. This setup will be referred to as "semi-supervised domain+task TL with x% of target data labelled".

4 Results

Table 2 shows the results for the ICBHI source data, and Table 3 gives the results for the HFLungV1 source data.

4.1 Training from scratch

The upper block of Tables 2 and 3 gives the AUROCs when training after random weight initialisation (He uniform). For Compass data, it can be seen that models trained from scratch using the target data performed best, but are inconsistent (high standard deviation). Models trained on the public data (source) and evaluated on the in-house Compass (target) data performed worse, indicating that a domain and/or task shift exists between these datasets.

For ZOL target data, training from scratch on ICBHI source data gives a significant[1] improvement ($p < 0.01$) compared to training on the ZOL target data. The differences in AUROC for wheezing between both source datasets and the target dataset are smaller for the ZOL data than for Compass data. This will be further detailed in the Discussion section.

4.2 Domain+task TL

The second block of Tables 2 and 3 shows the results of the domain+task TL experiment. In this experiment, models were first pre-trained using public source data and were then fine-tuned on the in-house target data. This domain+task TL could be considered as an upper bound for what is possible through TL.

Compared to training from scratch on the Compass target data, there is no improvement for crackles, and only a small improvement from $74.99 \pm 9.57\%$ to $75.74 \pm 6.63\%$ for wheezes (not significant) when fine-tuning from ICBHI data. When first training on HFLungV1 source data and later fine-tuning on Compass target data, no significant improvement occurs when compared to directly training on Compass target data.

When comparing training from scratch on the ZOL target data and domain+task TL from HFLungV1, a significant improvement in wheezing AUROC ($p < 0.005$) occurs. For ICBHI source data, no improvement can be seen.

[1] Two-sided Wilcoxon signed-rank test

Table 2. Area under the ROC curve (AUROC) for ICBHI source data and both in-house target datasets. All results are mean ± sample standard deviation across 15 runs.

ICBHI source data	**Compass, Leave-One-Patient-Out**		**ZOL, Test Set**
TL Method	**AUROC crackles (%)**	**AUROC wheezes (%)**	**AUROC wheezes (%)**
From scratch on in-house data (target)	69.60 ± 12.99	74.99 ± 9.57	82.20 ± 1.33
From scratch on ICBHI public data (source)	53.59 ± 7.66	40.77 ± 11.53	84.08 ± 1.78
Domain+task TL from source to target	69.13 ± 10.65	75.74 ± 6.63	82.74 ± 1.47
Semi-superv. domain+task TL, 75% of target data	67.12 ± 7.78	73.74 ± 4.59	83.95 ± 1.72
Semi-superv. domain+task TL, 50% of target data	66.71 ± 6.17	73.01 ± 4.25	83.40 ± 1.56
Semi-superv. domain+task TL, 25% of target data	65.58 ± 5.27	70.79 ± 5.28	83.35 ± 1.50
Domain TL from source to target data	49.88 ± 6.02	46.91 ± 9.53	83.75 ± 2.15

4.3 Domain TL

The lowest block of Tables 2 and 3 shows the results of the domain TL experiment. For domain TL from ICBHI to Compass, the AUROC for crackles decreases significantly ($p < 0.05$) from 53.59 ± 7.66% to 49.88 ± 6.02%. The AUROC for wheezes significantly improves ($p < 0.025$) from 40.77 ± 11.53% to 46.91 ± 9.53%. For HFLungV1 to Compass, both classes improve (not significant for crackles, $p < 0.05$ for wheezes).

For ZOL data, there is no improvement in wheezing AUROC for both source datasets.

4.4 Semi-supervised domain+task TL with subset of labelled target data

The third block of Tables 2 and 3 gives the results when the unsupervised DANN is used as an initialisation and the classifier is fine-tuned using a subset of labelled target data.

For ICBHI source data, both crackle and wheeze AUROCs improve significantly ($p < 0.005$) when adding 25% of labelled Compass data. When adding an additional 25% of labelled Compass data, there is again a significant ($p < 0.025$) improvement in wheezing AUROC. When using HFLungV1 as the source data, the same findings hold.

For ZOL target data, there is no improvement in adding a portion of labelled target data when building further from ICBHI source data. When starting from

Table 3. Area under the ROC curve (AUROC) for HFLungV1 source data and both in-house target datasets. All results are mean ± sample standard deviation across 15 runs.

HFLungV1 source data	**Compass, Leave-One-Patient-Out**		**ZOL, Test Set**
TL Method	**AUROC crackles (%)**	**AUROC wheezes (%)**	**AUROC wheezes (%)**
From scratch on in-house data (target)	69.60 ± 12.99	74.99 ± 9.57	82.20 ± 1.33
From scratch on HFLungV1 public data (source)	48.20 ± 6.50	38.74 ± 8.54	65.45 ± 1.95
Domain+task TL from source to target	68.25 ± 7.84	74.03 ± 6.64	83.72 ± 1.06
Semi-superv. domain+task TL, 75% of target data	63.89 ± 7.80	67.69 ± 9.29	80.83 ± 2.55
Semi-superv. domain+task TL, 50% of target data	63.62 ± 8.02	67.01 ± 10.03	80.66 ± 2.78
Semi-superv. domain+task TL, 25% of target data	60.49 ± 7.49	62.27 ± 10.11	77.47 ± 2.92
Domain TL from source to target data	52.07 ± 6.20	45.80 ± 7.14	66.37 ± 5.53

the unlabelled DA on HFLungV1 source data, a significant improvement ($p < 0.005$) can be seen in wheezing AUROC. When adding an additional 25% of labelled ZOL data, again a significant improvement ($p < 0.005$) of 3.19% occurs.

5 Discussion

5.1 Training from scratch

The first block in Tables 2 and 3 shows the model performance when training from scratch starting from either source or target data. When training from scratch on Compass target data, it can be seen that the standard deviations are large, indicating that the models perform inconsistently. One possible explanation for this is the limited number of patients in this dataset, i.e. the training and validation sets always contained data of 4 patients, and the test set always had one unseen patient (leave-one-patient-out test set). This inconsistency indicates the need for more training data, which leads to TL.

When training from scratch on source data and evaluating on Compass target data, the classification performance is bad. A possible explanation is that the ThinkLabs One stethoscope used to record the Compass data is not present in any of the other datasets, as indicated in Table 1. This could then be solved by adapting to the new domain. A second possible explanation could be the difference in the definition of the "crackle" class, i.e. Compass data also included rales and crepitations while the other data only included crackles. This requires

adapting the model to cope with the new task, which can only be achieved by (semi-supervised) domain+task TL.

For ZOL target data, the wheezing AUROC increases when training on ICBHI source data. One possible explanation is that the increased amount of training data helps, and that the presence of multiple stethoscopes in ICBHI data acts as an augmentation which helps create robustness. For HFLungV1 source data, the wheezing AUROC (65.45 ± 1.95%) is lower for the ZOL target data compared to training on ICBHI source data (84.08 ± 1.78%). Unfortunately, it is hard to find possible causes for this, as the majority of HFLungV1 data (261 of 279 patients) lack details about the patient population. For instance, HFLungV1 does contain data of mechanically ventilated patients, which is not present in any of the other datasets [4]. Another possible explanation could be that the learnt features differ much from the features learnt on ICBHI data.

5.2 Domain+task TL

The second block in Tables 2 and 3 gives the AUROCs when first training the model on source data, and then fine-tuning the model on target data. This approach can be interpreted as an upper bound in model performance (best case).

When first training on ICBHI source data and then fine-tuning on Compass target data, the classification performance increases. This could indicate that the model adapts to the new task (different pathologies and redefinition of crackle class). When comparing the results of domain+task TL against training from scratch on target data, the standard deviations decrease. This indicates that the models are more consistent after domain+task TL and could confirm that training from scratch on Compass data is difficult due to the limited number of patients. The same story holds when first training with HFLungV1 source data.

When starting from HFLungV1 and fine-tuning on ZOL target data, an improvement in wheezing AUROC occurs. Therefore, it is possible that a domain shift exists between HFLungV1 and ZOL datasets. The wheezing AUROC does not improve after domain+task TL from ICBHI to ZOL. A possible explanation is that the model overfits on the ZOL training data, due to the small dataset size.

5.3 Domain TL

The bottom block in Tables 2 and 3 shows the model performance when first training the model on source data, and then applying unsupervised DANN using both source and target data for the domain discriminator. By using DANN, the feature distributions of source and target data should align better. When compared to training from scratch on source data, domain TL often helps increase model performance. However, for Compass target data performance on crackle classification decreases from 53.59 ± 7.66% to 49.88 ± 6.02%. One possible explanation for this decrease is that Compass data also contains other discontinuous adventitious sounds (rales and crepitations), which were grouped together with

crackles. Since these do not occur in ICBHI, the feature extractor should be learnt using labelled target data to include these (task adaptation).

For ZOL target data, wheezing AUROC decreases from 84.08 ± 1.78% to 83.75 ± 2.15% (ICBHI) and from 65.45 ± 1.95% to 66.37 ± 5.53% (HFLungV1), indicating that the model "unlearns" features that were learnt from the source data.

5.4 Semi-supervised domain+task TL with subset of labelled target data

The third block in Tables 2 and 3 gives the model's classification performance when fine-tuning the classifier after domain TL. The feature extractor's weights were kept constant. For both source datasets, it can be seen that using 25% of labelled target data already results in an improved classification performance. The benefit of using labelled target data, is that the model can also adapt to a new task, whereas only applying domain TL can only adapt to new domains.

When doing semi-supervised domain+task TL from ICBHI to ZOL data, no improvement can be seen. This shows that there is only a domain shift when moving from ICBHI to ZOL data. This domain shift can be tackled by the domain TL.

6 Conclusion

This work evaluated domain+task TL, domain TL and semi-supervised domain+task TL in the context of lung sound event classification. First, CNN models were trained from scratch using either source or target data. Second, domain+task TL was applied by pre-training the CNN models on public lung sound data and fine-tuning on our in-house data. This resulted in an upper bound for classification performance, but requires the entire target dataset to be labelled. Third, domain TL was performed using unsupervised Domain Adversarial Neural Network (DANN). Since domain TL cannot adapt to the new task, the resulting models were not suitable for deployment. Fourth, semi-supervised domain+task TL was tested by fine-tuning the classifier using a portion of labelled target data. Adding 25% of labelled target data already increased the classification performance.

Future work could investigate whether incorporating additional context, e.g. the type of stethoscope or the auscultation position, into the model input can enhance performance without necessitating changes to the model parameters.

Acknowledgements. The authors would like to thank Flanders Innovation & Entrepreneurship Agency (VLAIO) (PlugNPatch, 3E230186) and KU Leuven IOF (COMPASS, 3H230337) for providing research funding.
The authors would also like to thank Future Health research group for annotating the ZOL data.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bohadana, A., Izbicki, G., Kraman, S.S.: Fundamentals of lung auscultation. New England Journal of Medicine **370**(8), 744–751 (2014). https://doi.org/10.1056/NEJMra1302901, https://www.nejm.org/doi/full/10.1056/NEJMra1302901
2. Clevert, D.A., Unterthiner, T., Hochreiter, S.: Fast and accurate deep network learning by exponential linear units (elus) (2016), https://arxiv.org/abs/1511.07289
3. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of Machine Learning Research **17**(59), 1–35 (2016), https://jmlr.org/papers/v17/15-239.html
4. Hsu, F.S., Huang, S.R., Huang, C.W., Huang, C.J., Cheng, Y.R., Chen, C.C., Hsiao, J., Chen, C.W., Chen, L.C., Lai, Y.C., et al.: Benchmarking of eight recurrent neural network variants for breath phase and adventitious sound detection on a self-developed open-access lung sound database—hf_lung_v1. PLoS One **16**(7), e0254134 (2021)
5. Hsu, F.S., Huang, S.R., Su, C.F., Huang, C.W., Cheng, Y.R., Chen, C.C., Wu, C.Y., Chen, C.W., Lai, Y.C., Cheng, T.W., Lin, N.J., Tsai, W.L., Lu, C.S., Chen, C., Lai, F.: A dual-purpose deep learning model for auscultated lung and tracheal sound analysis based on mixed set training. Biomedical Signal Processing and Control **86**, 105222 (2023). https://doi.org/https://doi.org/10.1016/j.bspc.2023.105222, https://www.sciencedirect.com/science/article/pii/S1746809423006559
6. Huang, D., Wang, L., Lu, H., Wang, W.: A contrastive embedding-based domain adaptation method for lung sound recognition in children community-acquired pneumonia. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2023). https://doi.org/10.1109/ICASSP49357.2023.10096794
7. Jacobs, M., Vuegen, L., Verresen, T., Schouterden, M., Ruttens, D., Karsmakers, P.: Exploring model architectures for real-time lung sound event detection. In: Proceedings of the 33rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN) (2025)
8. Kim, J.W., Bae, S., Cho, W.Y., Lee, B., Jung, H.Y.: Stethoscope-guided supervised contrastive learning for cross-domain adaptation on respiratory sound classification. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1431–1435 (2024). https://doi.org/10.1109/ICASSP48485.2024.10447734
9. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (2019), https://arxiv.org/abs/1711.05101
10. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering **22**(10), 1345–1359 (2010). https://doi.org/10.1109/TKDE.2009.191
11. Rocha, B.M., Filos, D., Mendes, L., Serbes, G., Ulukaya, S., Kahya, Y.P., Jakovljevic, N., Turukalo, T.L., Vogiatzis, I.M., Perantoni, E., Kaimakamis, E., Natsiavas, P., Oliveira, A., Jácome, C., Marques, A., Maglaveras, N., Paiva, R.P., Chouvarda, I., de Carvalho, P.: An open access database for the evaluation of respiratory sound classification algorithms. Physiological Measurement **40**(3), 035001 (March 2019). https://doi.org/10.1088/1361-6579/ab03ea, https://bhichallenge.med.auth.gr/, accessed 30 July 2024.

12. Tefera, Y., Van Baelen, Q., Meire, M., Luca, S., Karsmakers, P.: Constraint-guided learning of data-driven health indicator models: An application on bearings. International Journal of Prognostics and Health Management **16**(2) (2025)