Significance of having a large sound dataset for pig cough classification

S. Upadhyaya[1,2,*], D. Berckmans[1], W. Desmet[4], W. Buyens[1], E. Vranken[1] and P. Karsmakers[2,3]
[1]*SoundTalks N.V, Interleuvenlaan 15/c, B-3001, Leuven, Belgium*
[2]*KU Leuven, Department of Computer Science, DTAI, Kleinhoefstraat 4, B-2440 Geel, Belgium*
[3]*Flanders Make, DTAI-FET, B-3001, Leuven, Belgium*
[4]*KU Leuven, Department of Mechanical Engineering, LMSD, Celestijnenlaan 300, Leuven, Belgium*
[*]Corresponding author: Sreenivasa Upadhyaya, sreenivasa.upadhyaya@soundtalks.com

Abstract

Continuous monitoring of the respiratory health status of pigs by means of automated detection of respiratory symptoms offer a solution to treat animals in an early stage of disease development, and consequently reduces the use of antimicrobials. Lately, pig cough classification for respiratory health monitoring using sound has caught the attention of the research community. This study examines the challenges involved in scaling the cough classification solution to newer farms with different acoustical properties. The results show that the models developed on a dataset captured in a certain acoustic condition fail to generalize across different farms with different acoustics. Experiments indicate that the performance can drop by almost 27 percentage points. The usage of a large sound dataset improves to model the realistic variations that are present in the data and avoids having a model bias for a certain acoustic condition. The model performance increases from 80% to 92.4% when samples size is increased from 10,000 to 400,000 samples. We conclude that the size and the acoustical diversity of the sound dataset plays a vital role in achieving a reliable cough classification model.

Keywords: cough analysis, sound event classification, deep learning, transfer learning.

Introduction

The global demand for human food is ever increasing, with limited resources available to grow, which calls for the need of efficient livestock farming practices. Precision Livestock Farming (PLF) is a modern farming approach that utilizes advanced technologies to monitor, manage and optimize livestock production systems with a high level of accuracy and precision (Berckmans D., 2014). Monitoring livestock health is the most important aspect in PLF. The most significant health concern for swine producers today is respiratory diseases (Brockmeier et al., 2002). These hamper the growth of the animal resulting in higher feed conversion rates and extra treatment costs (Holck et al., 2003). If the animals are not treated in time, it can be fatal. All these lead to significant economic loss, reduced welfare and overhead in barn maintenance.

In the context of monitoring respiratory health of pigs, measuring and analysing sounds in the barn is becoming increasingly popular. Besides the fact that microphones are contactless and relatively cheap, there is no need for a direct line of sight and thus enables large group of animals to be monitored with a single sensor.

Continuous monitoring of sounds in the barn facilitates early warning of a respiratory illness or an external factor causing the coughing. The early warning system reduces the monitoring burden of the farmer and generates alerts to intervene at an early stage of the illness. This reduces the excessive use of antibiotics (Rathkjen et al., 2022) and promotes targeted medication for the illness. Apart from reducing the medication costs, mortality is also greatly reduced (Vranken et al., 2022; Polson, 2022). Overall, the welfare of the animal is enhanced along with the quality and profitability of the production.

Cough is a reflex action to clear any irritants in the airways. Cough is a fundamental symptom in many respiratory illnesses. The presence of coughing gives an indication of a possible illness or adverse condition in the barn. Hence, the classification of cough sounds becomes a vital part of an early warning system for respiratory illness. A good cough classifier should be able to detect cough sounds precisely among other vocalizations and noisy sounds.

Early studies for cough classification (Guarino et al., 2004), used a feature vector with energy, time-derivate of energy and mean power spectral density and was compared to the reference cough set feature vector using dynamic time warping. In (Moshou et al., 2001), a hybrid classifier with a 2–class probabilistic neural network and a 4–class Multi-Layer Perceptron (MLP) was used to distinguish coughs and metal clanging from other sounds. (Vandermeulen et al., 2015) used features that have physical meaning and a simple decision tree classifier to obtain a scream classifier in production environments.

With the success of using deep learning methods for image classification applications, there has been interest in the research community to use it for Sound Event Classification (SEC) tasks (Yin et al.,2021, Song et al.,2022, Shen et al.,2022). Recent works (Yin et al.,2021, Song et al.,2022) show significant improvements in pig cough classification performance with the use of deep learning methods. Typically, a Convolutional Neural Network (CNN) based classifier is used to extract sound event features with the input presented in time-frequency representation known as spectrogram.

Previous works (Ferrari et al., 2008; Shen et al., 2022; Yin et al., 2022) used a labeled dataset that was collected from one or two farms. In practice, there exists a large variation in the characteristics of barn sounds due to the difference in breed, age, sex, barn machinery noise, barn ambient noise and acoustics of the barn. These differences significantly hamper the cough classification performance when trained on one specific farm and tested on a farm with differences in the above-mentioned characteristics. The models trained with this data perform well on the test set composed from data acquired in the same or a similar site. The main contribution of this work is the analysis that was carried out to study the effect of the size (and variability) of the data set on the robustness of the cough classifier model to changing acoustic conditions of the environment. Furthermore, the challenges of scaling pig cough classification systems in production environments, with a large, labeled audio dataset are explored.

The remainder of the paper is organized as follows. Section Materials and methods describes the dataset used in the experiments followed by Cough classification processing pipeline. Experimental results are discussed in Section Results and discussion. Finally, we present the conclusions in section Conclusions.

Materials and methods

Experimental data

The sound data used in this analysis was recorded from commercial barns. The audio data was recorded using a SoundTalks® sound monitoring device (SoundTalksN.V Belgium, 2022). This device is a commercial product which comes with a sensor system, that includes a microphone, relative humidity, and temperature sensor, and an associated web application to track the respiratory health status. During the data collection the sound monitoring device was used in research mode, which enabled storing raw sound files on a hard disk attached to the data collection hardware. The data was collected in 16-bit PCM format at a sampling frequency of 22,050 Hz. The monitors were hung from the ceiling and placed at a height of 2 meters from the floor. The number of monitors placed in a compartment was decided based on the dimension of the compartment and, generally a separation of 20 meters was maintained between two monitors. The data used was from finishing pens with pigs between the age of 4 to 26 weeks. The collected data were

comprising from farms located in Europe and USA. The dataset included batches during different parts of the year.

Preselected audio events were labeled as cough or non-cough by trained labelers. Labelers were trained with multiple expert labeled cough examples. To ensure high quality of the labels, all events were labeled by multiple labelers and labelers were periodically evaluated. The dataset used belonged to 14 different farms in which the surface area of the rooms varied between 50 m^2 and 1,500 m^2. These farms have different architectures and different technologies for feeding and ventilation, leading to different acoustical characteristics in the recorded events.

For this study, 1,096,012 labeled events were used to create a training set of 1,012,012 events and a test dataset of 84,000 events. The training dataset contains randomly selected events from the 14 different farms. The test dataset contains 3,000 coughs and 3,000 non-coughs from each of the 14 different farms, randomly selected from the complete labeled set. This makes the total test dataset to have 84,000 events, with 42,000 cough and 42,000 non-cough sound events. The training and test set are exclusive such that no event is part of both the training and test set. The farms are anonymized and named as farm1, farm2 and so on until farm14.

Cough classification processing pipeline

The overview of the model architecture used in this work is shown in Figure 1. The raw audio waveform is converted into a time-frequency input representation (mel spectrogram). Next, a CNN based feature extractor derives meaningful feature maps. These feature maps are fed to the classifier, which is realized using a group of Fully connected Layers (FL). The output of the classifier is a binary class with class 0 representing non-coughs and class 1 representing coughs.

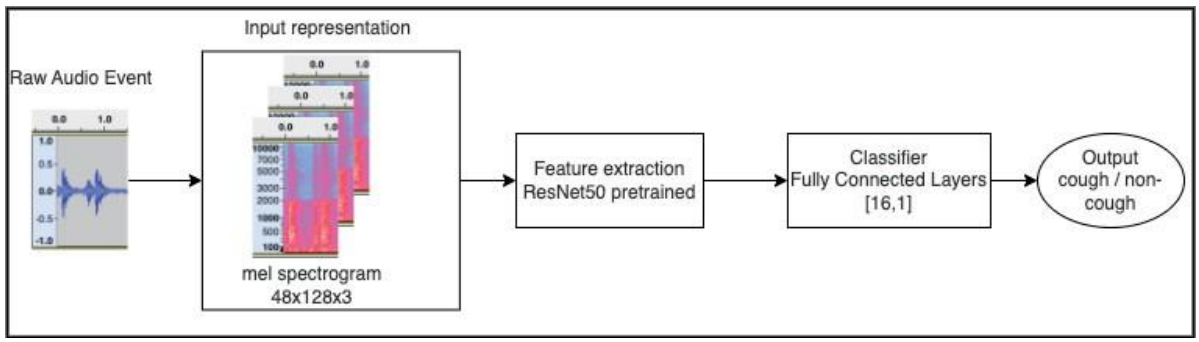

Figure 1: CNN based cough classification model.

Input representation

The input representation used is a mel-spectrogram. Time domain audio signals are transformed to the frequency domain using Short Time Fourier Transform. Subsequently, it is converted by several mel filter banks to a lower dimensional mel representation (Stevens et al., 1937). Mel is a perceptual scale for the audio frequencies. Mel spectral features are a popular choice for input representation in deep learning based acoustic event detection (Serizel et al., 2018). Conversion from time domain audio signal to mel-spectrogram was done using the mel-spectrogram function in librosa (McFee et al., 2015). Fmin, the lowest frequency to be considered to derive the mel spectral bins, was set to 50 Hz to avoid very low frequency bias due to the ventilation noise in the farms. STFT window length was set to 1024 with a hop length of 441 samples (20 ms). The number of mel-spectrogram bins was set to 128. This implies that every 20 ms of raw audio samples

generates a mel-spectrogram of length 128. Each raw audio event is converted to a mel-spectrogram of size 48×128 where 48 represents the number of time frames (of 20ms) and 128 corresponds to the number of mel frequency bins. The number of frames was set to 48 empirically, making it sufficient to contain the cough event in the window considered for classification. Figure 2 shows a typical cough event and its corresponding mel-spectrogram representation.

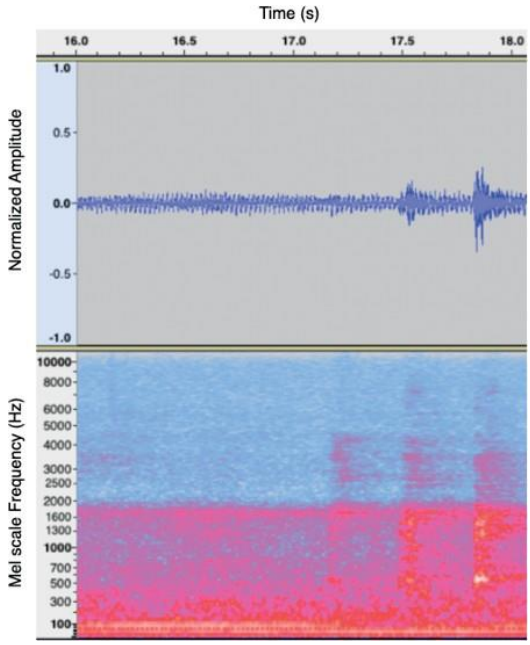

Figure 2: Sample representation of pig cough events: Time domain (above) and mel-spectrogram (below).

Model

The model architecture used in this work is shown in Figure 1. We will refer to this model as the baseline model in the remainder of the work. The core of this model is ResNet50 (He et al., 2016) CNN feature extractor model. Transfer learning (Zhuang et al., 2020) approach is used here by pre-training the model on Imagenet dataset (Russakovsky et al., 2015). Among the popular Imagenet pre-trained models like VGG, AlexNet, DenseNet and GoogleNet, ResNets seems to perform better in SEC tasks (Pahar et al., 2021). The CNN model is fine-tuned using the cough dataset. The mel-spectrogram generated from each audio event is repeated 3 times to convert it to a 3-channel input, making it compatible for the pre-trained ResNet50 model. The fully connected layers of the ResNet50 are dropped and replaced by a shallow fully connected layer of 16 neurons followed by an output layer of 1 neuron corresponding to the binary output class. The fully connected layer is trained along with the fine tuning of the CNN layers. The concept of residual training was first presented in (He et al., 2016). ResNet architecture was proposed for image classification tasks in ILSVRC (Russakovsky et al., 2015) and COCO 2015 (Lin et al., 2014) competitions and showed advantages in both complexity and performance compared to well-known deep learning model like VGG (Simonyan et al., 2014). The presence of skip connections makes the ResNet architecture resilient against vanishing gradients and helps the network layers to learn the identity functions efficiently. The skip connections feed the output of a layer to multiple subsequent layers. ResNets work by learning the residual functions with respect to the inputs at a given layer. The ResNet50 architecture has a layer depth of 50.

The cough classification model is trained with the Stochastic Gradient Decent (SGD) optimizer, with an initial learning rate of 1e-2. The training is done on mini batches of size 512. The model is trained for 50 epochs, which was sufficient for the model to achieve convergence. Binary cross entropy is used as the loss function. The training and experimentation are done on a server with AMD EPYC 7402P processor and NVIDIA GeForce RTX 2080 Ti GPUs. The models were implemented in Python3 using the TensorFlow/Keras deep learning framework.

Evaluation criteria

Typically, the performance of the binary audio event classification system is measured in terms of accuracy, precision, recall and f1-score. These metrics are derived from the values of True Positives (TP), True Negatives (TN), False Positives (FP) and False Neg- atives (FN)(see Equation 1-4). In this work, we used F1-score (Adavanne et al., 2018) as the metric, which jointly represents both Precision and Recall of the classifier system as a single metric.

$$Accuracy = (TP + TN)/(T\,P\ +\ T\,N\ +\ F\,P\ +\ F\,N) \quad (1)$$

$$Precision = TP/(T\,P\ +F\,P) \quad (2)$$

$$Recall = TP/(T\,P\ +\ F\,N) \quad (3)$$

$$F1 - score = (2 * Precision * Recall)/(Precision + Recall) \quad (4)$$

Results and discussion

Experiment 1

To understand the degree of variation present in the dataset, we trained multiple variants of the cough classification model. The only difference between these model variants is the difference in the training data. In the first experiment, we trained 14 models, each with the training data from one specific farm. For all the model variants, the performance analysis was done on the same test set as described in the Section Experimental data. This test data contains data from all the 14 farms. For each of the 14 model variants, we obtain 1 value for performance metric for that specific farm and 13 values for the 'left-out' farms. This is represented in Figure 3 where the dots represent the 'farm-specific' performance and the box plots represent the 'left-out' performance, which is the performance of all other farm-specific models on a given farm.

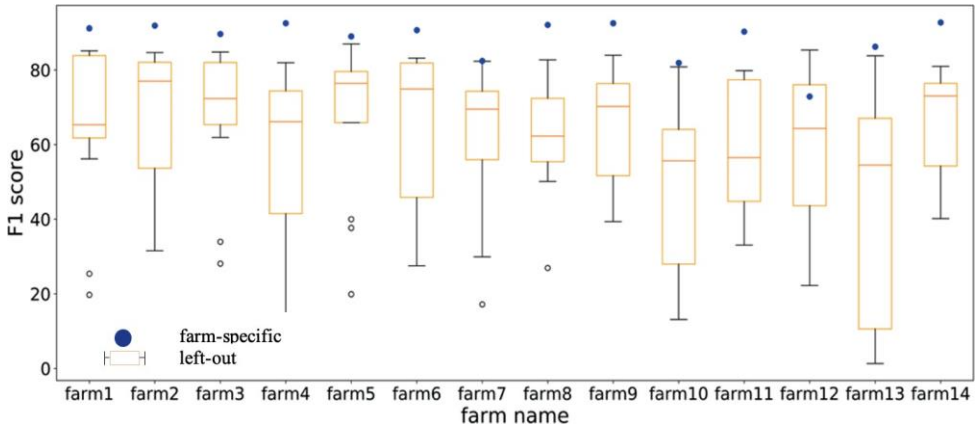

Figure 3: Farm wise F1-score for the 'farm-specific' and 'left-out' cases.

The mean performance in the 'left-out' case drops by more than 27 percentage points compared to the 'farm-specific' performance. Also, the standard deviation of the performance across farms increases. This points out the fact that the coughs and other sound events encountered across farms have diverse characteristics. Using the data from a single farm is not sufficient to model the variations in sound events. The models fit well in the 'farm-specific' case and fail to generalize well to "left-out" farms.

Experiment 2

In this experiment, we studied the significance of the size of the training dataset in the context of model generalization. The presence of the large dataset with more than 1 million labeled events gave the leverage to perform this experiment. Different baseline model variants were generated by training the baseline model with different sizes of the training dataset. We started with a training set of 10,000 events and kept increasing the training set size in steps of 10,000 events. From 100,000 events onwards, we increased in steps of 100,000 events until we reached the full training set size of 1 million events. The ratio of the samples from the farms as well as the ratio of coughs to non-coughs in the training dataset was maintained as it was in the original full training dataset.

The performance improvement with increasing training size is shown in Figure 4. The performance with 10,000 samples has a F1-score of 80%. We observe a sharp increase in the performance up to a training size of 100,000 samples. The performance increases linearly as the dataset size grows logarithmically. The performance significantly increases from 80% to 92.4% when samples size is increased from 10,000 to 400,000 samples. Post 700,000 samples, we hardly see any improvement using the current model.

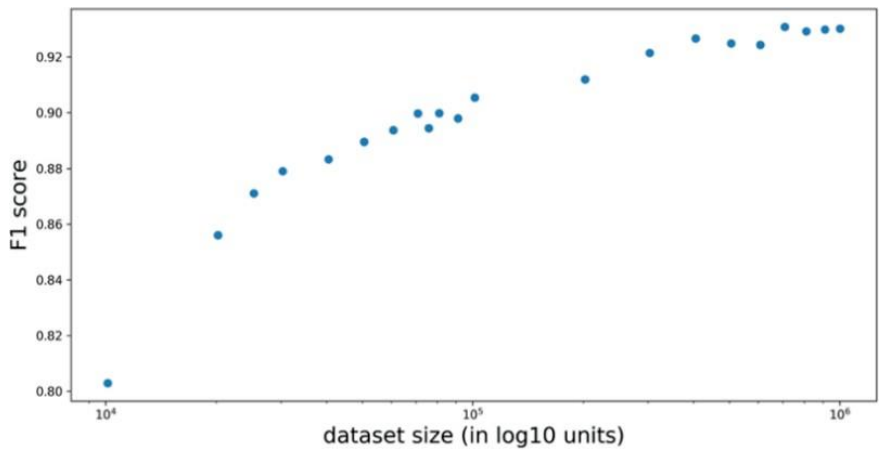

Figure 4: F1-score with logarithmically increasing dataset size.

Conclusions

To develop an accurate early warning system for pig respiratory illness, a robust and resilient cough classification system is a vital element. A deep learning based cough classifier has been effective in boosting the performance of such a system. Generalizing the model across different farms with varying acoustic conditions is very important. Experiment 1 showed the presence of variation present in the sound events across farms by analysing the results of models in farm-specific case or in cases where the farms are left-out from the training dataset. The vast difference of about 27 percentage points in the cough classification

performance calls for the need of an extensive dataset to cover the possible variations in the cough event characteristics.

Experiment 2 showed the significance of the large dataset to generalize the model. The models tend to overfit when trained on a dataset from few farms. The results in this experiment show a performance improvement with the logarithmic increase of the dataset size. The performance significantly increases from 80% to 92.4% when samples size is increased from 10,000 to 400,000 samples. The performance improves further but at a reduced rate when the training set size is further increased. It is expected that the significance of the data set size is even more important when more complex deep learning models are considered.

Acknowledgments

This work was supported by a Baekeland PhD grant of the Flanders Innovation and Entrepreneurship agency (VLAIO, Belgium) (HBC.2019.2216).

References

Adavanne, S., Politis, A., Nikunen, J., and Virtanen, T. (2018) Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing* 13(1), 34-48.

Berckmans, D. (2006) Automatic on-line monitoring of animals by precision livestock farming. *Livestock Production And Society* 287, 27-30.

Berckmans, D. (2014) Precision livestock farming technologies for welfare management in intensive livestock systems. *Revue Scientifique et Technique* 33(1), 189-196.

Brockmeier, S.L., Halbur, P.G., and Thacker, E.L. (2002) Porcine respiratory disease complex. *Polymicrobial Diseases*, 231-258.

Dale, P. (2022). Economic assessment of a sound-based precision livestock farming tool (SoundTalks) based on timing of intervention after a dual mycoplasma hyopneumoniae and PRRS virus seeder challenge in pigs. In: *Allen D. Leman Swine Conference*.

Ferrari, S., Silva, M., Guarino, M., Aerts, J.M., and Berckmans, D. (2008) Cough sound analysis to identify respiratory infection in pigs. *Computers and Electronics in Agriculture* 64(2), 318-325.

Guarino, M., and Costa, A. (2004) Automatic detection of infective pig coughing from continuous recording in field situations. *Rivista di Ingegneria Agraria (Italy)* 35, Issue 4.

He, K., Zhang, X., Ren, S., and Sun, J. (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Holck, J., and Polson, D. (2003) The financial impact of prrs virus. *The porcine Reproductive and Respiratory Syndrome Compendium. 2nd edition National Pork Board*, 51-58.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C.L. (2014) Microsoft coco: common objects in context. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings,* Part V 13.

Mao, A., Giraudet, C.S., Liu, K., De Almeida Nolasco, I., Xie, Z., Xie, Z., Gao, Y., Theobald, J., Bhatta, D., and Stewart, R. (2022) Automated identification of chicken distress vocalizations using deep learning models. *Journal of the Royal Society Interface* 19(191), 20210921.

McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E., and Nieto, O. (2015) Librosa: audio and music signal analysis in python. In: *Proceedings of the 14th python in science conference*.

Moshou, D., Chedad, A., Van Hirtum, A., De Baerdemaeker, J., Berckmans, D., and Ramon, H. (2001) An intelligent alarm for early detection of swine epidemics based on neural networks. *Transactions of the ASAE* 44(1), 167.

Pahar, M., Klopper, M., Warren, R., and Niesler, T. (2021) Covid-19 cough classification using machine learning and global smartphone recordings. *Computers in Biology and Medicine* 135, 104572.

Rathkjen, P.H., Jensen, M., Kragge, A., and Alonso, C. (2022). Sound based health monitoring performed by SoundTalks® significantly reduces the overall antibiotic consumption nursery facilities. In: *13th European Symposium of Porcine Health Management.*

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., and Bernstein, M. (2015) Imagenet large scale visual recognition challenge. *International journal of computer vision* 115, 211-252.

Serizel, R., Bisot, V., Essid, S., and Richard, G. (2018) Acoustic features for environmental sound analysis. *Computational analysis of sound scenes and events*, 71-101.

Shen, W., Ji, N., Yin, Y., Dai, B., Tu, D., Sun, B., Hou, H., Kou, S., and Zhao, Y. (2022) Fusion of acoustic and deep features for pig cough sound recognition. *Computers and Electronics in Agriculture* 197, 106994.

Simonyan, K., and Zisserman, A. (2014) Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv 1409, 1556*.

SoundTalks N.V, Belgium. (2022) *https://www.soundtalks.com/soundtalks/.* Last accessed December 2022.

Vandermeulen, J., Bahr, C., Tullo, E., Fontana, I., Ott, S., Kashiha, M., Guarino, M., Moons, C.P., Tuyttens, F.A., and Niewold, T. (2015) Discerning pig screams in production environments. *PLoS One* 10(4), e0123111.

Volkmann, J., Stevens, S., and Newman, E. (1937) A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America* 8(3), 208-208.

Vranken, E., Buyens, W., Ghysen, A., and Berckmans, D. (2022) The impact of respiratory health status on production losses in commercial pig farms. In: *European Conference on Precision Livestock Farming*.

Yin, Y., Tu, D., Shen, W., and Bao, J. (2021) Recognition of sick pig cough sounds based on convolutional neural network in field situations. *Information Processing in Agriculture* 8(3), 369-379.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2020) A comprehensive survey on transfer learning. In: *Proceedings of the IEEE* 109(1), 43-76.