ACTIVE LEARNING IN SOUND-BASED BEARING FAULT DETECTION

Maarten Meire^{1,2,3}, *Jeroen Zegers*^{4}, *Peter Karsmakers*^{1,2,3}

^{1} KU Leuven, Dept. of Computer Science, ADVISE-DTAI, Kleinhoefstraat 4, B-2440 Geel, Belgium
{maarten.meire, peter.karsmakers}@kuleuven.be
^{2} Leuven.AI - KU Leuven institute for AI
^{3} Flanders Make @ KU Leuven
^{4} Flanders Make vzw, CoreLab MotionS, 3001 Leuven, Belgium

ABSTRACT

Sound Event Classification (SEC) for fault detection of bearings in rotating machinery has recently shown good results. Bearing fault detection via microphones has advantages over the more traditional accelerometer-based solutions, in terms of ease of sensor deployment, non-intrusiveness and hardware cost. These novel SEC methods often use deep learning (DL), which require large amounts of labeled data. As events of faulty bearings are rare in practical scenarios, it can be time consuming to manually find and label examples of faults. Rather than labeling a complete dataset, active learning (AL) methods present the expert labeler with unlabeled samples that are expected to be the most informative in the learning process. This way the most interesting samples are labeled first, which allows to only annotate a subset of the dataset, while still retaining (close to) maximal accuracy. In this work a novel data set, that contains acoustic data from accelerated life time tests for bearings, is used to investigate the performance of two AL methods in terms of classification accuracy and number of additionally selected and annotated examples.

Index Terms— Active learning, Fault detection, Bearing monitoring, Transfer learning

1. INTRODUCTION

An important part in industrial applications is rotating machinery. Rolling Element Bearings (REB) are a common element in this machinery and most system failures can be attributed to these REB [1]. It is important to detect faults in these REB to prevent critical failures and this is most commonly done based on vibration analysis [2], however research has also been done towards using sound signals for REB fault detection [3]. Various data-driven approaches, usually Deep Learning (DL) based, have been investigated for the purpose of REB fault detection in the last few years. As these approaches are data-driven, and often based on DL, large amounts of data are required to train the associated models. For vibration analysis acquiring this data requires an accelerometer to be attached directly to the REB, which is not always trivial, especially in complex machines. By using a microphone this can be done without needing direct contact, making the data acquisition process easier, and it has been shown that using sound signals for fault detection is a promising alternative to vibration analysis [4, 5, 6, 7]. However, even if a lot of data can be acquired, the process of annotating this data remains time and cost intensive. Active Learning (AL) methods have been developed to reduce this cost by only annotating samples that are the most informative for learning algorithms [8, 9]. The focus of this work will be to use sound signals captured by a microphone in combination with AL methods for fault detection in REB.

In literature AL methods have already been used for fault detection in industrial applications. In [10] an extension of the entropy measure of model uncertainty was used to select the most informative samples to train a model that was learned on a data set with isolated and compound faults for REB fault detection. A combination of entropy and complexity was used in [11] to select samples for fault diagnosis in a gearbox showing a better performance using this combination. AL was applied to cellular networks in [12], with a comparison of 3 uncertainty based sampling methods, demonstrating their effectiveness. In [13] it is mentioned that using a single criterion strategy might not be stable and a new criterion is proposed that combines multiple commonly used criteria. A best versus second best uncertainty metric was used in [14] in combination with label propagation and ensembles to improve the performance of bearing fault diagnosis using a small training set.

The previously discussed works use vibration signals as data. However, as mentioned earlier, this work will focus on using sound signals. To the best of our knowledge, in the literature no prior work regarding AL for REB fault detection using sound is found. Nonetheless, AL in combination with sound has shown promising results in other fields. In [15] AL methods were evaluated using 2 synthetic sound event datasets for sound event detection and it was shown that training while keeping the original training set along with the annotated samples resulted in a better performance. A combination of AL and semi-supervised learning methods was used in [16] on a total of 3 datasets containing sound data for gender identification, speaker identification, and emotion detection. In [17] an alternating certainty sampling method was proposed where sometimes samples with high confidence were chosen instead of low confidence to improve the robustness against incorrect annotations. This method was evaluated on an urban sound dataset.

To compare AL methods a novel and unique accelerated bearing life time test dataset is used. It contains data captured using an accelerometer and 2 microphones. To the best of our knowledge, there is no public dataset that contains sound signals from bearing life time tests, as there are for vibration signals, e.g. IMS [18].

The rest of this paper is structured as follows. In Section 2 the AL methods that will be compared are explained. A detailed description of the experimental setup is given in Section 3, this includes the dataset, the preprocessing, the architecture and learning parameters of the models, and a description of the performed experiments. The results of the experiments are discussed in Section 4. Finally, conclusions and future work are given in Section 5.

2. ACTIVE LEARNING

In this section the (AL) methods that will be used in the experiments are discussed.

2.1. Uncertainty sampling

The first method, which is commonly used for AL, is based on selecting the samples for which the model predictions are most uncertain. For a classifier this often means that these samples are located close to the decision boundary. A simple yet commonly used metric to quantify predictive uncertainty is the information entropy,

$$H(X) = -\sum_{i=1}^{k} p_i log(p_i), \tag{1}$$

where X is the sample under evaluation, k is the amount of classes, and p_i is the estimated posterior probability for the $i-th$ class as predicted by a classifier. In this work a Convolutional Neural Network (CNN) is used to provide the probabilities for each class, the model will be described in Section 3.3.

2.2. Hybrid sampling

To avoid sampling multiple similar samples that have high prediction uncertainty the sampling criterion can be augmented with a novelty metric. The latter is referred to as hybrid sampling. In this way there is a potential to further decrease the annotation cost [8]. The Semi-Supervised Detection of Outliers (SSDO) [19] algorithm is used to calculate the novelty metric. This algorithm is based on k-means, but does not only take into account the distance to a cluster center, but also the size and relative position of the considered cluster. Formally, the hybrid sampling strategy is defined as follows:

$$S(X) = H(X) + \alpha N(X), \tag{2}$$

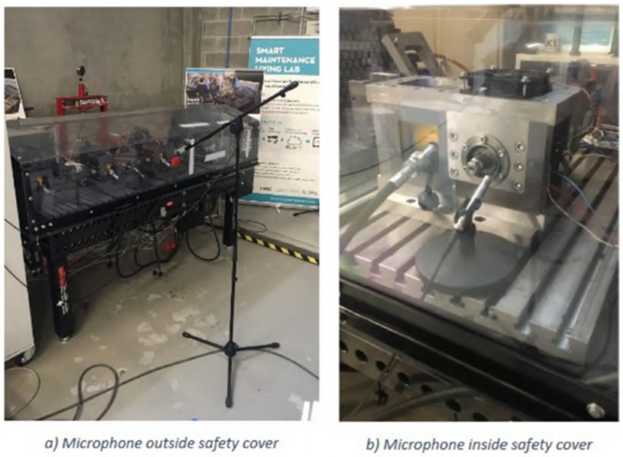
where $N(X)$ is the score obtained by SSDO for a sample X and α is a hyperparameter that balances both individual scores. Both scores have ranges $H(X) \in [0, 1]$ and $N(X) \in [0, 1]$. The sample X with the highest $S(X)$ is then selected as the most interesting.

3. EXPERIMENTAL SETUP

This section describes the dataset that was used in the experiments, as well as the preprocessing that was performed on the data, the model architecture and associated learning parameters that were used, and finally the performed experiments.

3.1. Dataset

This work uses a novel data set, collected by Flanders Make, that consists of data collected from multiple accelerated bearing life time tests. Data was collected using an accelerometer, a microphone inside the safety cover of a setup, and a microphone outside the cover, hereafter called the internal and external microphone, respectively. The setup can be seen in Figure 1, with the placement of both the internal and external microphone. The accelerometer was attached directly to the bearing housing. All sensors captured data with a sampling frequency of 50 kHz. A total of 64 accelerated life time tests, or runs, were performed. For each bearing a small indent was created on the inner races (IR) using a Rockwell-C indenter. The lifetime was further accelerated by applying a radial load of 9 kN. The test was stopped when either the stopping criteria of 20g peak vibrations was reached or the test had to be stopped due to safety concerns, e.g. overheating. The life time tests were performed with varying settings, e.g. fixed or varying rpm, and other setups running next to the test setup. It was also determined that the various life time tests not only resulted in inner race faults, but also in outer race and ball faults or in some cases it was considered as not faulty. The experiments in this work use a subset of this dataset, more specifically the data captured by the external microphone from only the runs with an inner race fault, with no additional running setups, and a fixed rpm during the run. Note that this rpm can vary between runs, e.g. 1800 rpm for one run and 2000 rpm for another. In total the considered subset contains 10 runs that match these criteria. The same setup was already used in previous works for accelerometer based fault detection [20].



a) Microphone outside safety cover

b) Microphone inside safety cover

Figure 1: The microphone setup used in the bearing life time tests.

In addition to the captured data, two sets of ground truth labelling are also provided. It should be noted here that this labelling is not based directly on the physical state of the bearing, but based on analysis of the data captured by either the accelerometer or the internal microphone. Using this labelling the moment in time p_f where the bearing starts having faulty behavior is determined. Data prior to p_f is then considered as healthy and data after p_f is considered faulty. As there are two sets of labelling, p_f is determined separately for each.

3.2. Preprocessing

In this work the raw audio data is first transformed to log mel spectra. This transformation is done using a window and hop size of 1s. A total of 64 mel filterbanks are then extracted. This leads to an input frame with shape (64,10), as each audio fragment is 10s long, which can then be passed to the models. As the data consists of multiple different runs, each run is separately standardized, using a running mean and standard deviation, to have, approximately, zero mean and unit variance. For the CNN the input frames are used directly, while for SSDO the mean and standard deviation for each filterbank are calculated over 10s and then stacked, resulting in a 128 dimensional feature vector.

After this preprocessing the data is split into 3 parts: 1) a training run that will be used to train an initial model, 2) a sampling set that will be used to sample points from, and 3) a test run that will be used to assess the generalization performance of the model. This split is done in a leave-one-run-out scheme, meaning that there will be 10 folds, as there are 10 available runs, with a single run in each test set. From the remaining 9 runs one is chosen as the training run and the other 8 are used as the sampling set. In this training run data is taken so that the amount of healthy and faulty samples is roughly equal. More specifically, this is done by taking all the data after p_f and taking the same amount of data directly prior to p_f. Then a maximum of 100s of data from the start of the run is also added. From the training run 20% will be used as validation.

3.3. Model architecture and learning parameters

The CNN model used in this work consists of 3 convolutional blocks, using 64, 64, and 32 filters, respectively, 2 fully connected blocks, both using 20 neurons, and a final fully connected layer as output. A single convolutional block is a sequence of a convolutional, batch normalization [21], maxpooling, and dropout [22] layer. The fully connected block contains the same sequence without the maxpooling. The leaky ReLu activation function was used for all layers, except the final layer, which uses a softmax activation. The filters in the convolutional layers are all size (7,7) and move with a stride of 1 in each direction. The maxpooling layers use a (2,2) window and move with a stride of 2 in each direction. All dropout layers use the default drop rate of 0.5.

A model was trained using the data described in Section 3.2 to serve as starting point for the AL methods. This model was trained for 100 epochs with the Adam optimizer [23] and a learning rate of $1e^{-3}$. If the validation loss did not improve for 10 consecutive epochs, the learning rate was halved. The weights were further regularized by the L_2 norm with a factor (λ) of $5e^{-6}$. All hyperparameters were empirically tuned independently from the test data.

The SSDO model, used for the hybrid sampling, was fitted using a contamination factor of $1e - 3$, meaning that 0.1% of the training data is considered as novel. The amount of clusters used by the algorithm is set to 5% of the amount of training data.

3.4. Experiments

In this work a comparison of AL methods will be made to investigate the model performance on an independent test set in terms of the employed number of additionally annotated examples. For this purpose, the same experiment was repeated twice, once with labels based on information from the internal microphone (Y_{Mic}) and once with labels based on the accelerometer (Y_{Acc}).

3 methods were compared to each other: 1) random sampling, where samples are chosen at random to annotate, this will serve as a baseline method , 2) uncertainty sampling, which uses the prediction probabilities of the CNN to determine what samples to annotate, as described in Section 2.1, and 3) hybrid sampling, which further incorporates a novelty metric, as described in Section 2.2. To evaluate the methods, first an initial CNN model was trained on the data from a single run that is available in the training partition, as described in Section 3.3. Next a first sample is selected for annotation using the considered sampling strategy. After the annotation, the sample was added to the training set and the CNN model was updated for 20 epochs and, if the hybrid sampling method was being used, the SSDO method was refitted. This process was repeated 200 times for each sampling method.

To quantify the performance the F1 score was used as a metric. The faulty class is considered to have a positive label. As the leave-one-run-out scheme was used, the mean and standard deviation of this metric are computed over the folds. However, it was noticed that the standard deviation was similar across the results, ranging from 0.15 to 0.2, hence it will not be shown for reasons of clarity.

4. RESULTS

In this section first the results of the experiment where the Y_{Mic} were used will be discussed. Thereafter, the same experiment is repeated but this time with Y_{Acc}. For both experiments the F1 score in terms of the additionally annotated samples, ranging from 0 to 200, will be tracked. As a baseline, the F1 score of a model trained using the full set of training samples (on average around 15000 samples) for 100 epochs was added.

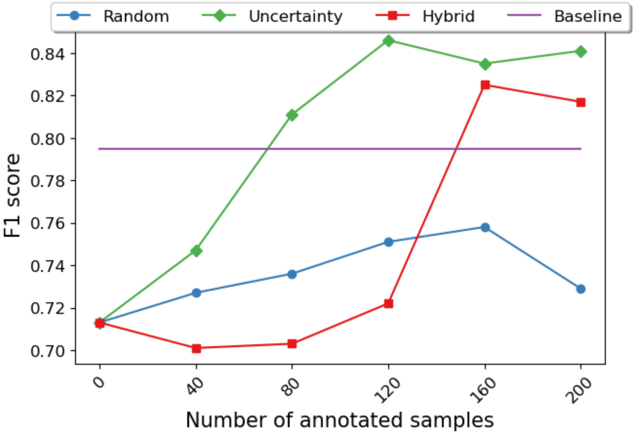


Figure 2: The F1 score comparison of the considered AL methods and the baseline on the test set in steps of 40 annotated samples using Y_{Mic}.

4.1. Microphone labels

The F1 scores attained by the CNN model on the test set using the AL methods described earlier and Y_{Mic} are shown in Figure 2. It can be seen that, while hybrid sampling does not perform as well with low amounts of annotated samples, both uncertainty and hybrid sampling outperform both random sampling and the baseline when respectively 80 and 160 samples are additionally annotated. This indicates that by using AL the amount of samples that need to be annotated can be significantly reduced, in this work by around 75 times or more. The difference between random sampling and uncertainty and hybrid sampling can likely be attributed to the sample selection. By inspecting these samples it can be seen that data around p_f for the various runs in the sampling set is chosen significantly more for uncertainty sampling, and to a lesser extent for hybrid sampling, while random sampling follows a more even distribution across the entire set, as is to be expected. This is empirically verified and will be discussed in Section 4.3. By choosing samples around p_f the model can learn a boundary between what is healthy and faulty. However, as there is a domain shift to the unknown bearing, it is expected that, while the boundary is likely to be improved, it will not be a perfect match. The lower performance of hybrid sampling up to 120 samples, could potentially be due to a smaller similarity between chosen samples increasing the complexity of the data in comparison to the other methods.

4.2. Accelerometer labels

The results attained on the test set when using Y_{Acc} are shown in Figure 3. It can be seen that these results are similar to the results attained using Y_{Mic}, with increasing F1 scores when more samples are annotated. Uncertainty sampling also slightly surpasses the baseline with 80 annotated samples. However, it does stagnate, and even performs slightly worse, afterwards. The difference between the 3 methods is smaller compared to the labels based on the microphone, especially towards higher amounts of annotated samples. It can be seen that hybrid sampling once again performs worse with

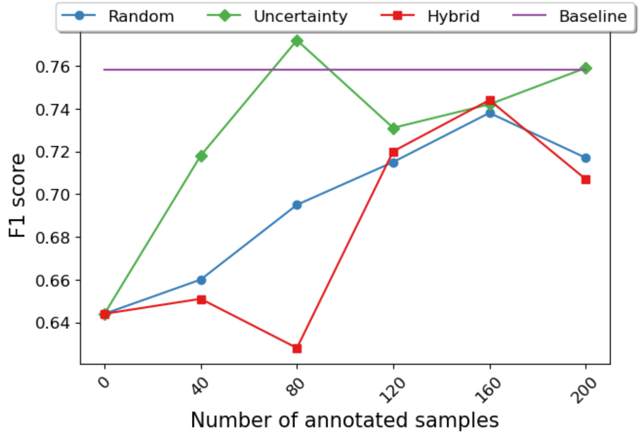


Figure 3: The F1 score comparison of the considered AL methods and the baseline on the test set in steps of 40 additionally annotated samples using Y_{Acc}.

up to 80 annotated samples. This is likely due to the same reasons as explained earlier. Additionally, as the accelerometer is expected to detect changes earlier, and thus p_f for Y_{Acc} is slightly earlier than p_f for Y_{Mcc}, the samples around p_f have an increased similarity, which will cause less of these samples to be chosen, as they will have a lower novelty metric. This possibly results in a too rough decision boundary with a worse performance as a consequence. The overall worse performance can likely also be attributed to the difference in p_f and corresponding labels. With Y_{Acc} the distribution of the healthy and faulty data is expected to overlap more. From the moment the accelerometer signals are starting to change the sound signals might still be very similar. This makes the problem more difficult with worse performance as a consequence.

4.3. Sample selection

As mentioned earlier, it was noticed that uncertainty, and to a lesser extent hybrid, sampling selected a significant amount of samples around p_f of a run. To illustrate this, an experiment was performed where only a single run was available in the sampling set and 50 samples, selected using uncertainty and hybrid sampling, were annotated following the same process as described in Section 3.4. The log mel spectrum of the specific run and a histogram of the selected samples is shown in Figure 4. It can be seen that uncertainty sampling selected 28 samples within 100 samples of p_f while 20 were selected, within the same group of samples, by hybrid sampling, indicating that indeed a significant amount of samples are selected around p_f by uncertainty, and to a lesser extent hybrid, sampling. Furthermore, it is indicated that samples are also selected with a larger selection by hybrid sampling, around noise events that pop-up in the healthy data, e.g. around 5000s or 500 samples, which would cause the model to learn the data is healthy, regardless of the disturbing noise events. This could have also contributed to the improved performance of AL compared to random sampling.

5. CONCLUSION AND FUTURE WORK

In this work we compared two AL methods, more specifically uncertainty and hybrid sampling, and a random sampling baseline method to evaluate the performance with regards to generalization to an unknown bearing when additional samples from known bearings are annotated. This was done using a novel dataset that contains accelerated bearing life time tests with data captured from an

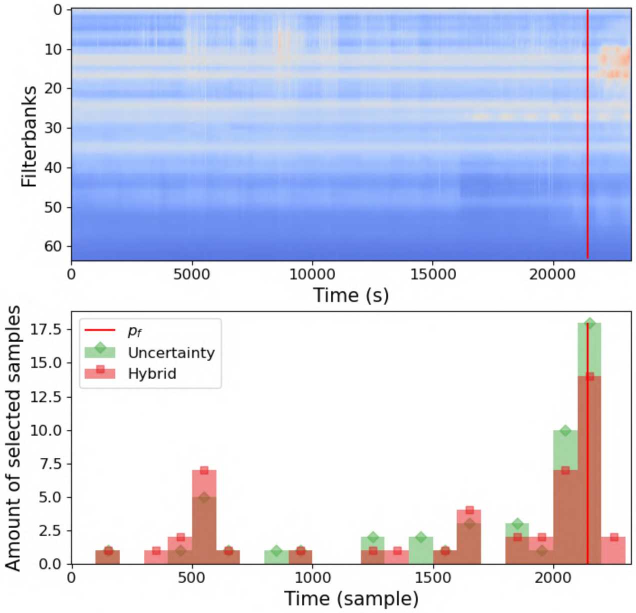


Figure 4: The log mel spectrum of the sampling run (top) and a histogram, with bins of 100 samples, of the selected samples (bottom) by uncertainty and hybrid sampling. The red line indicates p_f.

accelerometer, a microphone inside a safety cover, and a microphone outside the safety cover. Labels were provided based on the microphone inside the safety cover and based on the accelerometer.

It is indicated that, for the labels based on the microphone inside the safety cover, both AL methods outperform random sampling and also outperform the baseline that uses all data. Furthermore, the uncertainty sampling method does show a better performance compared to hybrid sampling. This is likely due to more samples being selected in the close vicinity of p_f. When looking at results for the labels based on the accelerometer, the difference between the methods is not as clear. However, uncertainty sampling still shows the best performance, also attaining a score similar to the baseline. The hybrid sampling method does not perform as well, likely due to smaller novelty metric between points near p_f, as the accelerometer can detect the fault earlier than the microphone. The results on both sets of labels indicate that it is possible to attain a similar, or better, performance to a method that does not use AL, while the amount of annotated samples was reduced by a factor of around 75.

In future research we will include different faults, e.g. outer race faults, into the experiments. We will also investigate the combination of label propagation with the AL methods.

6. ACKNOWLEDGMENT

The authors would like to thank the ACMON ICON project for funding the research presented in this paper. This research has received support of Flanders Make, the strategic research centre for the manufacturing industry. This research received funding from the Flemish Government (AI Research Program).

7. REFERENCES

[1] A. Nabhan, N. Ghazaly, A. Samy, and M. M.O, “Bearing fault detection techniques - a review,” *Turkish Journal of Engineering, Sciences and Technology*, vol. 3, 01 2015.

[2] D.-T. Hoang and H.-J. Kang, “A survey on deep learning based bearing fault diagnosis,” *Neurocomputing*, vol. 335, pp. 327 – 335, 2019.

[3] M. Altaf, M. Uzair, M. Naeem, A. Ahmad, S. Badshah, J. A. Shah, and A. Anjum, “Automatic and efficient fault detection in rotating machinery using sound signals,” *Acoustics Australia*, vol. 47, pp. 125–139, 2019.

[4] J. Pacheco-Chérrez, J. A. Fortoul-Díaz, F. Cortés-Santacruz, L. María Aloso-Valerdi, and D. I. Ibarra-Zarate, “Bearing fault detection with vibration and acoustic signals: Comparison among different machine leaning classification methods,” *Engineering Failure Analysis*, vol. 139, p. 106515, 2022.

[5] M. Iqbal and A. K. Madan, “CNC machine-bearing fault detection based on convolutional neural network using vibration and acoustic signal,” *Journal of Vibration Engineering &amp Technologies*, vol. 10, no. 5, pp. 1613–1621, Mar. 2022.

[6] J. Grebenik, Y. Zhang, C. Bingham, and S. Srivastava, “Roller element bearing acoustic fault detection using smartphone and consumer microphones comparing with vibration techniques,” in *2016 17th International Conference on Mechatronics - Mechatronika (ME)*, 2016, pp. 1–7.

[7] D. Xiao, C. Qin, H. Yu, Y. Huang, C. Liu, and J. Zhang, “Unsupervised machine fault diagnosis for noisy domain adaptation using marginal denoising autoencoder based on acoustic signals,” *Measurement*, vol. 176, p. 109186, 2021.

[8] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, B. B. Gupta, X. Chen, and X. Wang, “A survey of deep active learning,” *ACM Comput. Surv.*, vol. 54, no. 9, oct 2021.

[9] E. Lughofer, “Hybrid active learning for reducing the annotation effort of operators in classification systems,” *Pattern Recognition*, vol. 45, no. 2, pp. 884–896, 2012.

[10] Y. Jin, C. Qin, Y. Huang, and C. Liu, “Actual bearing compound fault diagnosis based on active learning and decoupling attentional residual network,” *Measurement*, vol. 173, p. 108500, 2021.

[11] J. Chen, D. Zhou, Z. Guo, J. Lin, C. Lyu, and C. Lu, “An active learning method based on uncertainty and complexity for gearbox fault diagnosis,” *IEEE Access*, vol. 7, pp. 9022–9031, 2019.

[12] M. Chen, K. Zhu, R. Wang, and D. Niyato, “Active learning-based fault diagnosis in self-organizing cellular networks,” *IEEE Communications Letters*, vol. 24, no. 8, pp. 1734–1737, 2020.

[13] Z. Liu, J. Zhang, X. He, Q. Zhang, G. Sun, and D. Zhou, “Fault diagnosis of rotating machinery with limited expert interaction: A multicriteria active learning approach based on broad learning system,” *IEEE Transactions on Control Systems Technology*, vol. 31, no. 2, pp. 953–960, 2023.

[14] C. Jian, K. Yang, and Y. Ao, “Industrial fault diagnosis based on active learning and semi-supervised learning using small training set,” *Engineering Applications of Artificial Intelligence*, vol. 104, p. 104365, 2021.

[15] Z. Shuyang, T. Heittola, and T. Virtanen, “Active learning for sound event detection,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2895–2905, 2020.

[16] S. Karlos, C. Aridas, V. G. Kanas, and S. Kotsiantis, “Classification of acoustical signals by combining active learning strategies with semi-supervised learning schemes,” *Neural Computing and Applications*, vol. 35, no. 1, pp. 3–20, Feb. 2021.

[17] Y. Wang, A. E. Mendez Mendez, M. Cartwright, and J. P. Bello, “Active learning for efficient audio annotation and classification with a large amount of unlabeled data,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 880–884.

[18] H. Qiu, J. Lee, J. Lin, and G. Yu, “Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics,” *Journal of Sound and Vibration*, vol. 289, pp. 1066–1090, 02 2006.

[19] V. Vercruyssen, W. Meert, G. Verbruggen, K. Maes, R. Bäumer, and J. Davis, “Semi-supervised anomaly detection with an application to water analytics,” in *2018 IEEE International Conference on Data Mining (ICDM)*, 2018, pp. 527–536.

[20] T. Ooijevaar, K. Pichler, Y. Di, S. Devos, B. Volckaert, S. V. Hoecke, and C. Hesch, “Smart machine maintenance enabled by a condition monitoring living lab,” in *8th IFAC Symposium on Mechatronic Systems MECHATRONICS 2019*, vol. 52, no. 15. IFAC, 2019, pp. 376–381.

[21] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *CoRR*, vol. abs/1502.03167, 2015.

[22] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[23] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *International Conference on Learning Representations*, 12 2014.