Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 225 (2023) 1073–1082

Procedia Computer Science

www.elsevier.com/locate/procedia

27th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2023)

Context-aware Acoustic Signal Processing

Liane-Marina Meßmer[a,b,*], Christoph Reich[a], Djaffar Ould Abdeslam[b]

[a]*Institute for Data Science, Cloud Computing and IT-Security (IDACUS), Hochschule Furtwangen - University of Applied Sciences, Robert-Gerwig-Platz 1, 78120 Furtwangen im Schwarzwald, Germany; l.messmer@hs-furtwangen.de*

[b]*Institut de Recherche en Informatique, Mathématiques, Automatique et Signal (IRIMAS) - Université de Haute Alsace, 61 rue Albert Camus, 68093 Mulhouse, France; djaffar.ould-abdeslam@uha.fr*

Abstract

Data processed in context is more meaningful, easier to understand and has higher information content, hence it derives its semantic meaning from the surrounding context. Even in the field of acoustic signal processing. In this work, a Deep Learning based approach using Ensemble Neural Networks to integrate context into a learning system is presented. For this purpose, different use cases are considered and the method is demonstrated using acoustic signal processing of machine sound data for valves, pumps and slide rails. Mel-spectrograms are used to train convolutional neural networks in order to analyse acoustic data using image processing techniques.

Keywords: Context-aware Deep Learning; Ensemble Neural Networks; Xception; Acoustic Signal Processing; Mel-Spectrograms

1. Introduction

A particular situation can be described by any information which characterises it, like a surrounding framework for objects, places, things or media. This kind of information is called context [21], [3]. Even in the field of (AI)-systems added context becomes increasingly more important for predictions made by such systems. The results of contextual models are much more detailed, accurate and more valuable, because some user or domain specific information is included in the knowledge of the system [5]. It is difficult to integrate context directly into a neural network, it is easier to integrate it into the system as a pre- or post processing step, therefore the use of Ensemble Neural Networks is one possibility to integrate context into a deep learning system [10].

An Ensemble Neural Network architecture is a modular system, which splits of a big deep learning model into several smaller ones which are acting as an ensemble network [38]. Each ensemble member has its own assignment and can be added or removed from an ensemble structure depending on the related context. The final prediction of such

* Corresponding author. Tel.: + (49) 07723 920-2383
E-mail address: l.messmer@hs-furtwangen.de

1877-0509
10.1016/j.procs.2023.10.095

a network consists of multiple separate predictions, which found consensus instead of one single prediction made by a neural network. In the pre-processing stage, some of the networks can be selected, and in the post-processing step, the results can be selected, aggregated and chosen according to the related context. With this approach it is possible to create a context-aware learning system, by using different ensemble members for different problems [11].

Acoustic signals obtain information about a signal source and the related environment including background noise. The prerequisite for a hearing system is the processing of acoustic data [6]. In advanced digital signal processing, high-performance computers are used to extract information from acoustic signals to create, for example, an analysing Artificial Intelligence (AI) system [13].

Hearing systems can be used in a wide variety of applications, for example to classify biodiversity in bird sounds and to detect which birds are singing in an acoustical signal, containing multiple bird sounds with background noise [37]. Furthermore in the area of underwater signal detection e.g. for biodiversity monitoring in coral reefs with the use of underwater ocean sounds [17]. Another use case for acoustic signal processing is, for example, detection of anomalous machine sounds in industrial environments, to detect if a machine is defect of not [28]. Digital signals can even be used in healthcare applications, for example to detect abnormal heart sounds [31] or in the area of the care for dementia patients. For example, familiar everyday or animal sounds can be meaningful in reminiscence sessions, when working with People with Dementia (PwD) [15],[14]. Acoustic signals, which are useful to evoke memories in PwD, should preferably not contain any other noises or background noises, which can be examined by processing such signals.

This work demonstrates acoustic signal processing using machine sound monitoring as an example use case. In the industry, unplanned machine breakdown leads to high costs and delays in a company's production and profits [34]. With the fourth industrial revolution, digital signals and information's are captured, monitored, exchanged and can be analysed via AI systems [33]. Every machine on a factory floor produces audible sounds, which can be captured, digitised and analysed [32]. In order to decrease machine downtime, this paper presents an AI-based audio classification approach, to investigate which machine is running on a production floor and to detect whether a machine is at a standstill or not.

The literature has shown, that it is useful to process acoustic data, by using a two-dimensional time-frequency representation, in order to train Convolutional Neural Networks (CNN) with visual acoustic data [23]. The use of spectrograms is common in the area of deep acoustic signal processing. Since the frequency resolution of a mel-spectrogram [9] is higher than that of a spectrogram, the experiments in this work are carried out with mel-spectrograms [39]. In addition, mel-spectrograms performed well in other experiments [30].

In this paper, a neural network based ensemble approach is presented, by using visual acoustic machine signals. The goals are to find out whether it is possible to integrate context into an acoustic signal processing system, using an ensemble of convolutional neural networks and to find out whether it makes sense to choose a context-sensitive approach or whether this approach is even necessary to solve the described problem. In addition, the work investigates if machine monitoring to detect and reduce machine downtimes by using CNNs and mel-spectrograms is possible.

This paper consists of 6 Sections. Section 2 deals with related work. The dataset and data preprocessing is described in Section 3. In Section 4 the training process is presented and the results are compared and evaluated, additionally some observations are described. Finally, the conclusion and future work follow in Section 5.

2. Related Work

Nascimento et al. [27] presented a context-aware machine learning approach for autonomous street light control and air quality prediction. Instead of using ensemble neural networks for context integration, they select a specific single neural network by using contextual information. Aljulayfi and Djemame [10] present an approach to integrate context into a machine learning environment by using a self adaptive system to support the autoscaling decision in Edge Computing. Their goal is to choose the machine learning algorithm depending on a different context. For the experiment, they used Support Vector Machines, Linear Regression and Neural Networks. The difference between these two works and ours is that we evaluate an acoustic file using several fixed networks in parallel and aggregate the results. We integrate the context into our system in a different way.

Bai et al. [2] created an Xception based method for bird sound recognition, with the goal to detect multiple bird species out of an acoustic bird signal. Just like in our work, they used an imaging procedure and trained a Xception

network with the help of transfer learning. For training, they used mel-spectrograms and, unlike us, spectrograms as well. The work proposed by Müller et al. [22], presents an approach for acoustic malfunctioning machine detection, also with the use of transfer learning. Unlike our work, they do not use any ensemble techniques, but similar to our work is the signal processing and learning technique. The has shown, that the use of an pre-trained neural network achieved better results. Their training samples consists out of mel-spectrograms generated from files of the MIMII dataset, similar to our work. But an other difference is, that they use ResNets for training in combination with different anomaly detection models.

There are a lot of works and use cases using ensemble neural networks in the literature. For example in breast cancer detection [7], weather forecasting [18], bio-image classification [25] or to classify skin lesions [12]. Some more works existing in the area of ensemble learning to process acoustic signal data, for example in healthcare to detect abnormal heart sounds from patients [31], to classify animal audio data [24] as it could be usable for the treatment of dementia patients or in the area of traffic noise prediction [29].

Nanni et al. [26] created an ensemble based audio classification approach. They used sound files from bird calls, cat sounds and environmental sounds. Similar to our work, they applied mel-spectrograms for visual audio representations to train Convolutional Neural Networks. They conclude, that the use of different neural network architectures maximised their ensemble performance. Unlike their work, we use a dataset-based ensemble instead of an ensemble of different network architectures. Ahmed et al. [1] presented a sound based machine failure detection system with the use of ensemble neural networks. They used the MIMII dataset and a dataset called "ToyAdDMOS" to create visual signal representations. Different to our work, they have not used mel-spectrograms as training inputs, instead they utilised MFCCs, Chroma Vectors or Spectral Entropys. Just like in our work, they created a neural network ensemble with different charges of their dataset. Our work sets itself apart by focusing on machine monitoring rather than machine failure detection.

3. Dataset

The experiments are based on the "MIMII Dataset: Sound Dataset vor Malfunctioning Industrial Machine Investigation and Inspection", created and published by Purohit et al. [32]. With this dataset, the currently existing gap is filled in terms of public machine audio classification datasets, to assist the research in machine-learning and signal-processing. The dataset contains audio files from four different machine types: valves, fans, slide rails and pumps. For each type of machine, different operating modes are captured. The signals of the valve were recorded with different opening and closing times, while the sounds of the pump were recorded during suction and discharge into a water pool. The acoustic of the fan was recorded under normal operation and that of the slide rail with slide repetitions at different speeds. In addition there are audio files recorded from well working machines and from defect machines, to resemble real factory failures. To simulate a factory under real life conditions, background noise from multiple real factories was added to the machine recordings. They provide the data with different levels (−6dB, 0dB, 6dB) of Signal to Noise Ratio (SNR).

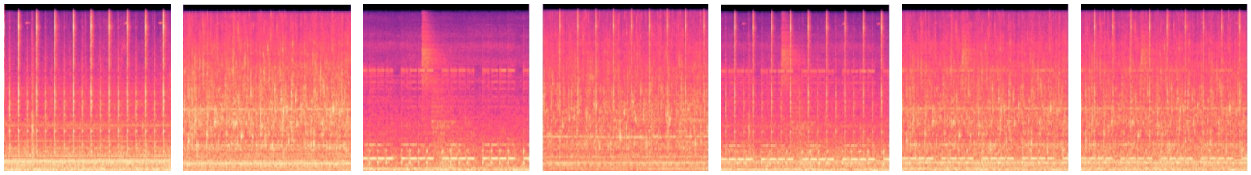
The recordings are related to seven different machine types of one machine class (e.g. six types of valves, pumps, fans and slide rails). In total there are 26092 audio segments for normal conditions and 6065 for anomalous conditions. Each audio segment has a length of 10s and was recorded as 16-bit audio signal sampled at 16 kHz in a reverberated environment. Note, that in this publication only sound files for well working machines are used, because it deals with acoustic machine monitoring and not with machine fault detection or predictive maintenance.

3.1. Preprocessing

In a typical shopfloor scenario, it is common for all machines to be operating simultaneously under normal conditions. The first goal is to be able to create a real factory environment with the given data, which is recorded from individual machine signals. It is possible for each machine to run alone, but also in all possible combinations with other machines. To simulate a real factory floor environment, we select three machine types of the dataset (valve, pump and slide rail (slider)) and superimposed the audio files in all possible combinations. Only audio files from three machines are used to perform the experiments, because the signal of the fan is the weakest of the machine

signals, which lead to an exclusion of this machine. The seven classes ”valve”, ”pump”, ”slider”, ”valve_pump”, ”valve_slider”, ”pump_slider”, ”valve_pump_slider” emerged from this step.

Since it has proven successful in the past to solve an audio-based problem with an image processing approach, the next step was to convert the audio files into mel-spectrograms [23]. Python offers a framework with which mel-spectrograms can be generated out of audio files. The framework is called Librosa [19]. With the help of the framework (Version 0.8.1 [20]), all seven audio classes are converted into mel-spectrograms. A spectrogram is generated with a maximum rate of 9000Hz, a window size of 2048, the ”hann” windowing function and a hop length of 512. Figure 1 shows an excerpt of the results of this converting step. Each picture is related to the first file of a machine class, with the label ”00”, extracted from the MIMII Dataset and with a SNR of 6dB.



(a) valve. (b) pump. (c) slider. (d) valve_pump. (e) valve_slider. (f) pump_slider. (g) valve_pump_slider.

Fig. 1. Mel-Spectrogram Representations.

To test if color images providing better training results than gray-scale images, we compared the training results of a color image dataset to the results of a gray-scale dataset. There are no significant differences. We decide to use gray-scale images in terms of training performance.

For training, we used only audio files from the first machine of a machine class, tagged with the label ”00”, from the MIMII dataset. The dataset has to be balanced in terms of the number of training examples per class. To enable optimal training of a network, the classes ”pump” and ”slider” were aligned to the class ”valve”, which has 991 images. This results in a dataset with 991 images per machine class.

Part of this work is to compare the outputs of a single neural network to the results of an ensembling network, with the aim to show if a context-aware network structure can be achieved or not and to figure out which approach gains better results. Hence, two datasets are needed, one, to train a Single Neural Network and one to train an Ensemble of Neural Networks.

3.2. Single Network Datasets

In case of training a single neural network, a distinction has to be made between two categories: One network is trained with the whole seven class dataset and one network is trained only on the three raw classes ”valve”, ”pump” and ”slider”, extracted from the dataset. In this case, the generated mixed class content is only used to test the training results.

When a single neural network is trained on all seven classes, the dataset has 991 images per class for each SNR (-6dB, 0dB, 6dB). This results in 20811 training samples, which are divided into a random 70/25/5 split, per class. The dataset is distributed as follows: There are 14595 images for training, 5187 for validation and 1029 for testing. This dataset is called ”Single Network Dataset 1”.

In the other case to train a network just on the raw class content, the raw training samples consists of 8919 images, 991 per class for each SNR and was also divided into a random 70/25/5 split, per class. The mixed class content consists out of the classes ”valve_pump”, ”valve_slider”, ”pump_slider” and ”valve_pump_slider”, also with 991 images per class for each SNR. This results in 11892 images. We will refer to this dataset as ”Single Network Dataset 2”.

In order to obtain a balanced test dataset needed for the comparison of the results, 5% of the images are randomly selected, so that the number of test images assigned to the ”valve”, ”pump” and ”slider” training classes is equal to the number of test images from the mixed class contents. Therefore, these classes are also used to test the network results. The final distribution of the dataset is as follows: The total number of training images is 6255, the total number of validation images is 2223 and the total number of test images is 1029.

3.3. Ensemble Network Datasets

Since the chosen ensemble strategy in this work is to ensemble on the dataset, in order to create multiple specialised networks, a separate dataset is needed for each specialised ensemble member. Therefore, the single network dataset has to be split differently in order to fulfill the problem definition. The network ensemble consists of three different participants, which are represented by the three classes ”valve”, ”pump” and ”slider”.

Each of the ensemble participants should only be able to decide whether it is the class in which they are specialised or not. Therefore, there are two labels per ensemble member, the specialised class ”valve”, ”pump” or ”slider” and ”other”. Unlike the single network dataset, mixed classes are now also used for training. The class ”other” is described by all classes in which the specialised class does not occur. The three ensemble members and the corresponding training classes are shown in Table 1, e.g. ensemble member ”valve” is trained on the classes ”valve” and ”other” and the class ”other” consists of training samples from ”pump”, ”slider” and ”pump_slider”. The classes ”valve_slider”, ”valve_pump and ”valve_pump_slider” are omitted in this example, as they contain the specialised class ”valve”. Therefore, they cannot appear in any other class, as they would otherwise distort the result set. The same applies to all other ensemble members.

Table 1. Distribution of the Training Classes from the Network Ensemble.

Ensemble Members	valve	pump	slider
Specialised Training Class	valve	pump	slider
Training Class: ”other”	pump	slider	valve
	slider	valve	pump
	pump_slider	valve_slider	valve_pump

Given that the class ”other” is composed of multiple training classes, it comprises a total of 8919 images. The specialised class, however, contains only 2973 images. To compensate for this imbalance, the images of the class ”other” were reduced so that they contain a total of 2970 images (evenly distributed among the classes that occur in the ”other” dataset class). Before splitting the data into the ensemble classes, we subjected the data to the same 70/25/5 split as in the single class datasets, so that the images that make up the test data are the same and a comparison of the results can be made, and so that the test data does not occur simultaneously in the training or validation data. The split results in 2085 training images per specialised class and 696 training samples per ”other” machine class, which results in a total of 2088 images. In total, there are 4173 training images. The same applies to the validation files. The class on which a network is specialised consists of 741 validation files and the class ”other” of 246 images, three times. In total, there are 738 validation images. For testing the ensemble networks, the same test dataset was used as for testing the network, which was trained with the single network datasets. This dataset contains −6db, 0dB and 6dB SNRs for training and testing. We will refer to it as ”Ensemble Dataset 1”.

To test with which noise intensity the results became weaker, we created two additional datasets containing different Signal to Noise ratios. The number of training samples in the different classes were reduced accordingly. One Dataset contains only 0dB and 6dB SNRs for training and −6dB, 0dB and 6dB SNRs for testing (”Ensemble Dataset 2”) and the other 0dB and 6dB SNRs for training and testing (”Ensemble Dataset 3”). Every ”Ensemble Dataset” is related to one specific machine class (valve, pump or slider) shown in Table 1. In total, there are 9 Ensemble Datasets used for training.

4. Evaluation and Comparison

This section describes and compares the results generated by training the single neural networks and by training the network ensemble.

4.1. Training

Literature has shown, that the use of transfer learning is useful in the area of CNN-based acoustic signal processing [16], [35]. Especially the Xception architecture generates promising results [2]. Hence, all networks in this publication are trained on a Xception [8] Network Architecture, pretrained on ImageNet, with Tensorflow and Keras framework. The networks are trained with 60 epochs and a batch size of 16. We use categorical cross-entropy as a loss function with adam optimizer and the standard learning rate of 0.001.

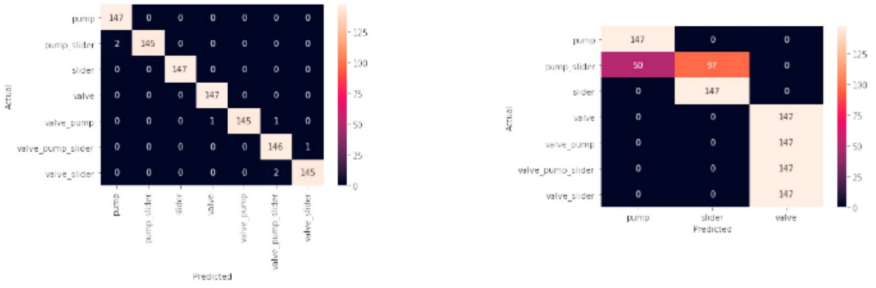
Training architectures are presented in Figure 2, with (a) depicting the architecture used for training a single network on "Single Network Dataset 1 and 2", while (b) showcases the ensemble network architecture trained on the "Ensemble Dataset 1, 2 and 3", for every type of machine. The result aggregation in (b) amounts to a simple concatenation of the sub predictions.

(a) Single Neural Network Architecture. (b) Ensemble Neural Network Architecture.

Fig. 2. Training Architectures.

4.2. Results Single Neural Networks

The single neural network was trained on the Single Network Dataset 1 and 2. The results are presented in the form of a confusion matrix in Figure 3: (a) shows the result generated with the Single Network Dataset 1 and (b) shows the results generated with the Single Network Dataset 2.



(a) Training Results "Single Network Dataset 1". (b) Training Results "Single Network Dataset 2".

Fig. 3. Training Results Single Neural Network.

Results in Figure 3 (a) show promising inference results for classifying all classes, including mixed classes (Single Network Dataset 1), only the mixed classes have misclassifications, but these account for a normal, rather low rate of misclassified results. However, the number of training classes increases overproportional as the number of machines on a shopfloor rises, leading to impractically long training and inference times.

To address this issue, a new dataset using only atomic classes "valve", "pump" and "slider" was created for training, with all mixed classes used for inference. Results in Figure 3 (b) demonstrate perfect classification of atomic classes, like in (a), but mixed classes cannot be detected with one inference pass. You only see for example, that the signal of the class "valve" is the strongest, since it was detected by all mixed classes containing this class. Furthermore you can see, that the signal of the "pump" seems to be less concise compared to the signal of the "slider", since the class "slider" has more right predictions than the class "pump" related to the mixed class classification.

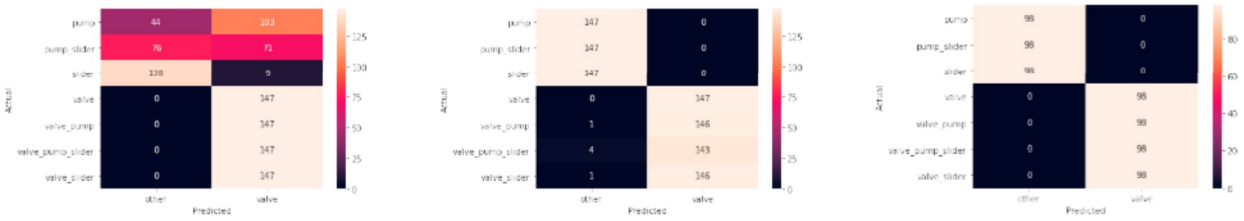
To achieve multiple class recognition of mixed audio files without the increasing number of training classes, a context-aware system using an ensemble neural network approach is proposed. Each ensemble member is trained to recognize a single machine from a signal with multiple machines, enabling simple addition (without network re-training) of new ensemble members when new machines are added.

4.3. Results Ensemble Neural Networks

To bring context into a machine learning system, we created a network ensemble and trained each ensemble member on a different sub-problem. For each machine type ”valve”, ”pump” and ”slider”, three networks are trained (Ensemble Network Dataset 1-3).

The results gained with this strategy are shown in Figure 4, Figure 5 and Figure 6. Figure 4 contains all results of the ensemble member ”valve”, Figure 5 contains all results of the ensemble member ”pump” and Figure 6 contains all results generated by the ensemble member ”slider”. In each Figure 3-5 (a) describes the results generated by training and testing with −6dB, 0dB and 6dB SNRs (Ensemble Dataset 1), (b) describes the results gained by training with 0dB and 6dB SNRs, but testing with all −6dB, 0dB, 6dB SNRs (Ensemble Dataset 2), (c) describes the results of the ensemble member achieved by training and testing with 0dB and 6dB SNRs (Ensemble Dataset 3). The comparison between (a), (b) and (c) is necessary, to find out from which signal to noise ratio the results become worse.

So, first let’s look at the results of the ensemble member valve, which are presented in Figure 4.



(a) Training, Testing: -6dB, 0dB, 6dB. (b) Training: 0dB, 6dB; Testing: -6dB, 0dB, 6dB. (c) Training, Testing: 0dB, 6dB.

Fig. 4. Training Results of the Ensemble Member ”Valve”.

As we have already seen from the Single Neural Network experiment, the valve’s signal is the strongest. The results of the training with the ensemble dataset confirm this observation. If we look at (c), we can see that the signal from the valve was recognised without an error, both from the atomic class and from the mixed classes. The classes that do not contain the valve signal were also classified as ”other” without error. Next, it is the turn of (b). Here, too, the strength of the concise signal of the valve can be seen. Training was done with a lower signal to noise ratio than testing. Nevertheless, the machine is recognised almost without outliers from the result set. The atomic class ”valve” was recognised without errors. ”valve_pump”, ”valve_pump_slider” and ”valve_slider” were detected incorrectly a total of six times, which amounts to a rather low, normal error rate. The last result we can look at is (a). This ensemble member was trained and tested including the lowest SNRs. The results obtained are partially worse than those generated in (b), which was trained without the lowest SNR, but tested with the same SNRs. However, this refers only to the classes which do not include the class ”valve”. The signal of the class ”pump” and the signal of the class ”pump_slider” were unusually often recognised as the signal of the class ”valve”, instead of the class ”other”. The atomic class slider was wrong classified 9 times (”valve” instead of ”other”), this deficit can be seen as a standard deviation. It remains to be seen whether the observation that the results from (a) are worse than those from (b) will be confirmed in the next attempts or whether this is a random event.

The next results we will look at are the results of the ensemble member pump, shown in Figure 5. As the name suggests, this ensemble member was trained on the ensemble dataset ”pump”. The two classes which should be recognised are ”pump” and ”other”. All experiments (a), (b) and (c) show, that the signal of the pump must be very weak because its hard to get it out of a mix of different machine signals. The best result of the classes ”valve”, ”valve_pump”, ”valve_pump_slider” and ”valve_slider” generated within (a) (training and inference with all SNRs), whereby the atomic signal ”pump” could be classified without error in all three cases. The biggest difference between the three result sets is in the two classes ”pump_slider” and ”slider”. In (a) 55% of the slider test samples were correctly classified as ”other”. In (b) 45% were correctly classified and in (c) 65%.

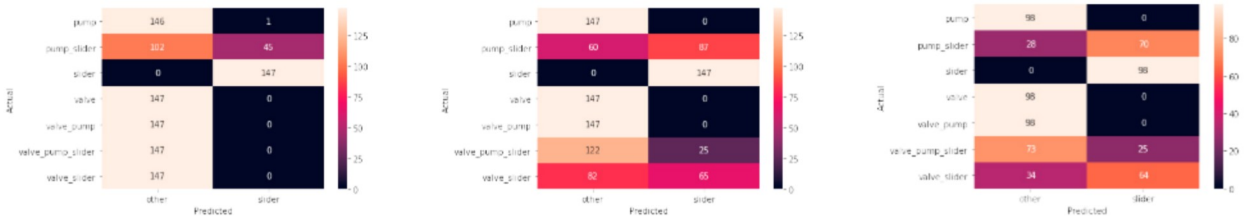
As the SNR decreases, the results deteriorate (difference between (a) and (b) and (c)). The reason that the correctness of the results of the Ensemble Member ”pump” is only about 50% is that the noise of the pump is the least significant in relation to the classes ”valve” and ”slider”. The background noise of the class ”slider” is probably partly equated with the noise of the pump. The last class in the result set that we look at is the class ”pump_slider”. In

(a) 63% were recognised correctly and in (b) 81% of the classifications were correct. In (c), 24% were recognised incorrectly (”other” instead of ”pump”) and 76% correctly. Again, the correctness of the results in (b) is higher than the correctness of the results in (a), which in turn supports the assumption that training with SNRs higher than zero increases the noise tolerance in the result set even more than training with higher noise data. Furthermore, it must be mentioned that the class ”pump_slider” is the only mixed class from which the pump is recognised. In all other mixed classes, which contain the class ”pump”, the machine class ”valve” dominates in such a way that the weak signal of the pump can no longer be detected.

(a) Training, Testing: -6dB, 0dB, 6dB. (b) Training: 0dB, 6dB; Testing: -6dB, 0dB, 6dB. (c) Training, Testing: 0dB, 6dB.

Fig. 5. Training Results of the Ensemble Member ”Pump”.

The final member of the ensemble that was trained is the slider. The training results are shown in the confusion matrix in Figure 6. The three atomic classes ”valve”, ”pump” and ”slider” were all classified almost one hundred percent correctly. Only once in (a) the class ”pump” was recognised as ”slider”. In addition, the class ”valve_pump” was classified as ”other”, which is completely correct, because the signal of the slider does not occur in this audio mix. Now the classes ”pump_slider”, ”valve_pump_slider” and ”valve_slider” remain. In (a) the class ”pump_slider” was 31% correct and 69% incorrect classified, in (b) 60% correct and to 40% incorrect and in (c) 72% correct and 28% incorrect. From these results it can be concluded that although the class pump is less visible in the mel-spectrogram, the slider probably resembles the noise of the pump to some extent, at least when they are mapped simultaneously on a mel-spectrogram. The class ”valve_pump_slider” produced more wrong than correct predictions in all three cases. This is due to the fact that the strong signal of the valve overlays all other signals. The same problem occurs in the class ”valve_slider”. Also in the case of this ensemble member, the phenomenon occurs that results from (a) are worse than the results of the training from (b).



(a) Training, Testing: -6dB, 0dB, 6dB. (b) Training: 0dB, 6dB; Testing: -6dB, 0dB, 6dB. (c) Training, Testing: 0dB, 6dB.

Fig. 6. Training Results of the Ensemble Member ”Slider”.

4.4. Observations

Training with a higher signal to noise ratio, which means with less noise, produces better results than training with a lower signal to noise ratio, which means with more noise, related to the same result set. The prediction with respect to all atomic classes is very good and almost always one hundred percent correct. The class ”valve” is best recognised with respect to the complete set of test data due to its strong signal, which is visually reflected concisely in the mel-spectrogram. This is illustrated in Figure 1 (g). Despite the superimposed signals of all machines, ”valve” stands out the most.

Problems occur when recognising valve and pump, as the valve signal is mapped so strongly onto a mel-spectrogram that the pump signal is lost in it. This can be checked visually with the aid of Figure 1 (f), which shows a mixture of signals from the two machine classes, ”pump” and ”slider”. However, the signal of the slider is partially overlaid by the signal of the pump. In addition, the signal of the pump on the mel-spectrogram resembles general background noise. Therefore the phenomenon occurs that the pump is often confused with decreasing SNRs and the resulting higher background noise. Simultaneously, the signal of the slider is lost in that of the pump or it is superimposed by the background noise.

Despite the visual similarity of some classes, the predictions are in general very good. There is only a need for improvement in a few places for the use of the technology on a real factory floor.

5. Conclusion and Future Work

This work describes an approach, on how to integrate context with the use of Ensemble Neural Networks into a deep learning system, by using visual machine signal features as mel-spectrograms. The MIMII dataset was used with Xception Networks for training. With the proposed approach it is possible to use an Ensemble Neural Network architecture to integrate context into an AI-based learning system, for visual acoustic signal processing. Furthermore, with our use case, it is necessary to use such an approach, because of the over-proportional rising of training classes with an increasing number of machines on a factory floor. In addition the results of a single neural network can be improved with this approach. One more benefit is, that the network is easy to extend and adaptable to another factory hall.

In the future, other diagrams describing acoustic singals can be used. For example the Mel Frequency Cepstral Coefficience (MFCC) diagram oder a Chromagram [4]. In this work the evaluation and concatenation of the result subsets is made by humans, later a specific ensembling method can be used. Here, the bagging and stacking methods would lend themselves particularly well, as we have performed a dataset bootstrap [11]. Weaker signals being overlaid by resembling background noise could be improved, by using different denoising algorithms, e.g. a Kalman filter [36].

Acknowledgements

This work is part of the project ”Digital Technologies for the Care of People with Dementia (DIDEM)”, which was made possible by funding from the Carl Zeiss Foundation.

References

[1] Ahmed, F., Nguyen, P., Courville, A., 2020. An ensemble approach for detecting machine failure from sound, in: Proceedings of the 5th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE), Tokyo, Japan, pp. 2–4.
[2] Bai, J., Chen, C., Chen, J., 2020. Xception based method for bird sound recognition of birdclef 2020.
[3] Baldauf, M., Dustdar, S., Rosenberg, F., 2007. A survey on context-aware systems. International Journal of Ad Hoc and Ubiquitous Computing 2, 263–277.
[4] Bhatia, R., Srivastava, S., Bhatia, V., Singh, M., 2018. Analysis of audio features for music representation, in: 2018 7th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO), pp. 261–266. doi:10.1109/ICRITO.2018.8748783.
[5] Brézillon, P., 1999. Context in artificial intelligence: I. a survey of the literature. Comput. Artif. Intell. 18, 321–340.
[6] Candy, J., 2008. Signal processing in acoustics: Science or science fiction? Acoustics Today 4. doi:10.1121/1.2994726.
[7] Chennamsetty, S.S., Safwan, M., Alex, V., 2018. Classification of breast cancer histology image using ensemble of pre-trained neural networks, in: Campilho, A., Karray, F., ter Haar Romeny, B. (Eds.), Image Analysis and Recognition, Springer International Publishing. pp. 804–811.
[8] Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1251–1258.
[9] Combrinck, H., Botha, E., 1996. On the mel-scaled cepstrum. department of Electrical and Electronic Engineering, University of Pretoria .
[10] Fouopi, P., Srinivas, G., Knake-Langhorst, S., Köster, F., 2016. Object detection based on deep learning and context information.
[11] Ganaie, M., Hu, M., Malik, A., Tanveer, M., Suganthan, P., 2022. Ensemble deep learning: A review. Engineering Applications of Artificial Intelligence 115, 105151. URL: https://doi.org/10.1016%2Fj.engappai.2022.105151, doi:10.1016/j.engappai.2022.105151.
[12] Harangi, B., Baran, A., Hajdu, A., 2018. Classification of skin lesions using an ensemble of deep neural networks, in: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2575–2578. doi:10.1109/EMBC.2018.8512800.

[13] Hartmann, W.M., Candy, J.V., 2014. Acoustic signal processing. URL: https://doi.org/10.1007/978-1-4939-0755-7_14, doi:10.1007/978-1-4939-0755-7_14.
[14] Houben, M., Brankaert, R., Bakker, S., Kenning, G., Bongers, I., Eggen, B., 2020a. The role of everyday sounds in advanced dementia care, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. p. 1–14. URL: https://doi.org/10.1145/3313831.3376577, doi:10.1145/3313831.3376577.
[15] Houben, M., Brankaert, R., Kenning, G., Eggen, B., Bongers, I., 2020b. The perspectives of professional caregivers on implementing audio-based technology in residential dementia care. International journal of environmental research and public health 17, 6333.
[16] Koike, T., Qian, K., Kong, Q., Plumbley, M.D., Schuller, B.W., Yamamoto, Y., 2020. Audio for audio is better? an investigation on transfer learning models for heart sound classification, in: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 74–77. doi:10.1109/EMBC44109.2020.9175450.
[17] Lin, T.H., Akamatsu, T., Sinniger, F., Harii, S., 2021. Exploring coral reef biodiversity via underwater soundscapes. Biological Conservation 253, 108901. URL: https://www.sciencedirect.com/science/article/pii/S0006320720309599, doi:https://doi.org/10.1016/j.biocon.2020.108901.
[18] Maqsood, I., Khan, M., Abraham, A., 2004. An ensemble of neural networks for weather forecasting. Neural Computing and Applications 13, 112–122. doi:10.1007/s00521-004-0413-4.
[19] Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, Oriol Nieto, 2015. librosa: Audio and Music Signal Analysis in Python, in: Kathryn Huff, James Bergstra (Eds.), Proceedings of the 14th Python in Science Conference, pp. 18 – 24. doi:10.25080/Majora-7b98e3ed-003.
[20] McFee, B., Metsai, A., McVicar, M., Balke, S., Thomé, C., Raffel, C., Zalkow, F., Malek, A., Dana, Lee, K., Nieto, O., Ellis, D., Mason, J., Battenberg, E., Seyfarth, S., Yamamoto, R., viktorandreevichmorozov, Choi, K., Moore, J., Bittner, R., Hidaka, S., Wei, Z., nullmightybofo, Hereñú, D., Stöter, F.R., Friesch, P., Weiss, A., Vollrath, M., Kim, T., Thassilo, 2021. librosa/librosa: 0.8.1rc1. URL: https://doi.org/10.5281/zenodo.4782663, doi:10.5281/zenodo.4782663.
[21] Mehra, P., 2012. Context-aware computing: Beyond search and location-based services. IEEE Internet Computing - INTERNET 16, 12–16. doi:10.1109/MIC.2012.31.
[22] Müller, R., Ritz, F., Illium, S., Linnhoff-Popien, C., 2021. Acoustic anomaly detection for machine sounds based on image transfer learning. URL: https://doi.org/10.5220%2F0010185800490056, doi:10.5220/0010185800490056.
[23] Nam, J., Choi, K., Lee, J., Chou, S.Y., Yang, Y.H., 2019. Deep learning for audio-based music classification and tagging: Teaching computers to distinguish rock from bach. IEEE Signal Processing Magazine 36, 41–51. doi:10.1109/MSP.2018.2874383.
[24] Nanni, L., Costa, Y.M., Aguiar, R.L., Mangolin, R.B., Brahnam, S., Silla, C.N., 2020. Ensemble of convolutional neural networks to improve animal audio classification. EURASIP Journal on Audio, Speech, and Music Processing 2020, 1–14.
[25] Nanni, L., Ghidoni, S., Brahnam, S., 2021a. Ensemble of convolutional neural networks for bioimage classification. Applied Computing and Informatics 17, 19–35.
[26] Nanni, L., Maguolo, G., Brahnam, S., Paci, M., 2021b. An ensemble of convolutional neural networks for audio classification. Applied Sciences 11, 5796.
[27] Nascimento, N., Alencar, P., Lucena, C., Cowan, D., 2018. A context-aware machine learning-based approach, in: Proceedings of the 28th Annual International Conference on Computer Science and Software Engineering, pp. 40–47.
[28] Nishida, T., Dohi, K., Endo, T., Yamamoto, M., Kawaguchi, Y., 2022. Anomalous sound detection based on machine activity detection. arXiv:2204.07353.
[29] Nourani, V., Gökçekuş, H., Umar, I.K., 2020. Artificial intelligence based ensemble model for prediction of vehicular traffic noise. Environmental Research 180, 108852. URL: https://www.sciencedirect.com/science/article/pii/S0013935119306498, doi:https://doi.org/10.1016/j.envres.2019.108852.
[30] Pandey, S.K., Shekhawat, H.S., Prasanna, S.R.M., 2019. Deep learning techniques for speech emotion recognition: A review, in: 2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA), pp. 1–6. doi:10.1109/RADIOELEK.2019.8733432.
[31] Potes, C., Parvaneh, S., Rahman, A., Conroy, B., 2016. Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds, in: 2016 Computing in Cardiology Conference (CinC), pp. 621–624. doi:10.22489/CinC.2016.182-399.
[32] Purohit, H., Tanabe, R., Ichige, K., Endo, T., Nikaido, Y., Suefusa, K., Kawaguchi, Y., 2019. Mimii dataset: Sound dataset for malfunctioning industrial machine investigation and inspection. URL: https://arxiv.org/abs/1909.09347, doi:10.48550/ARXIV.1909.09347.
[33] Resende, C., Folgado, D., Oliveira, J., Franco, B., Moreira, W., Oliveira-Jr, A., Cavaleiro, A., Carvalho, R., 2021. Tip4. 0: industrial internet of things platform for predictive maintenance. Sensors 21, 4676.
[34] Roosefert Mohan, T., Preetha Roselyn, J., Annie Uthra, R., Devaraj, D., Umachandran, K., 2021. Intelligent machine learning based total productive maintenance approach for achieving zero downtime in industrial machinery. Computers & Industrial Engineering 157, 107267. URL: https://www.sciencedirect.com/science/article/pii/S0360835221001716, doi:https://doi.org/10.1016/j.cie.2021.107267.
[35] Tsalera, E., Papadakis, A., Samarakou, M., 2021. Comparison of pre-trained cnns for audio classification using transfer learning. Journal of Sensor and Actuator Networks 10, 72.
[36] Welch, G., 2014. Kalman filter. doi:10.1007/978-0-387-31439-6_716.
[37] Xie, J., Hu, K., Zhu, M., Yu, J., Zhu, Q., 2019. Investigation of different cnn-based models for improved bird sound classification. IEEE Access 7, 175353–175361. doi:10.1109/ACCESS.2019.2957572.
[38] Zhao, Y., Gao, J., Yang, X., 2005. A survey of neural network ensembles, in: 2005 International Conference on Neural Networks and Brain, pp. 438–442. doi:10.1109/ICNNB.2005.1614650.
[39] Ćirić, D., Perić, Z., Nikolić, J., Vučić, N., 2021. Audio signal mapping into spectrogram-based images for deep learning applications, in: 2021 20th International Symposium INFOTEH-JAHORINA (INFOTEH), pp. 1–6. doi:10.1109/INFOTEH51037.2021.9400698.