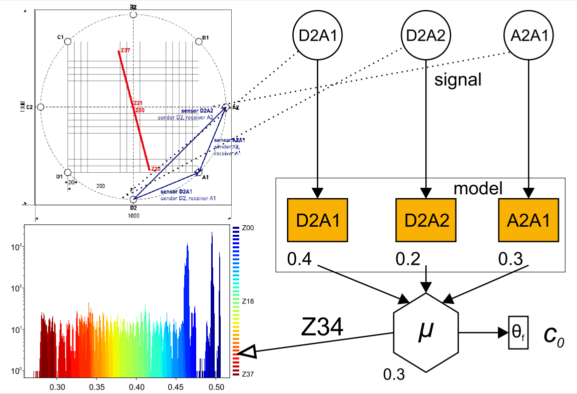
Machine Learning for Anomaly Assessment in Sensor Networks for NDT in Aerospace

Ivan Kraljevski, Frank Duckhorn, Constanze Tschöpe, and Matthias Wolff

***Abstract*—We investigated and compared various algorithms in machine learning for anomaly assessment with different feature analyses on ultrasonic signals recorded by sensor networks. The following methods were used and compared in anomaly detection modeling: hidden Markov models (HMM), support vector machines (SVM), isolation forest (IF), and reconstruction autoencoders (AEC). They were trained exclusively on sensor signals of the intact state of structures commonly used in various industries, like aerospace and automotive. The signals obtained on artificially introduced damage states were used for performance evaluation. Anomaly assessment was evaluated and compared using various classifiers and feature analysis methods. We introduced novel methodologies for two processes. The first was the dataset preparation with anomalies. The second was the detection and damage severity assessment utilizing the intact object state exclusively. The experiments proved that robust anomaly detection is practically feasible. We were able to train accurate classifiers which had a considerable safety margin. Precise quantitative analysis of damage severity will also be possible when calibration data become available during exploitation or by using expert knowledge.**



***Index Terms*— Machine learning, non-destructive testing, ultrasonic transducers.**

I. Introduction

THE expansion of the aerospace industry has led to an increase in the number of aging aircraft which are still in service [1]. Novel composite materials have been developed, which possess high strength-to-weight ratio and are used in aircraft, space vehicles and in other industrial applications. These materials are usually exposed to high loads and climatic factors progressively causing dangerous defects. Even the smallest flaws in the structure can lead to catastrophic failures. To ensure safety and airworthiness, it is necessary to employ new and innovative structural health assessment (SHA) techniques for fast and reliable inspection of aircraft parts [2].

Non-destructive testing (NDT) methods—e.g., based on acoustics, X-Rays, eddy current, or images—are capable of detecting defects during production as well as during usage. A recent and extensive review of NDT methods for defect detection and characterization in composites in aircraft structures is given in [3]. One of the most common SHA methods in the aerospace industry is the ultrasonic non-destructive testing [4]–[7].

In contrast to passive acoustic emission, ultrasonic NDT systems are actively employing transducers to send and receive ultrasonic waves to and from a specimen [8]. Typical excitation signals being sent to the specimen are wavelets or chirp signals with various center frequencies, allowing for the discovery of size and shape of extremely small damages and discontinuities (in the order of a signal wavelength) [9].

Earlier ultrasonic NDT systems mainly relied on the interpretation of readings by trained and experienced human operators. Later, the advances in signal processing and machine learning opened the possibility of building more sophisticated automated ultrasonic NDT systems with near-human detection performance [10].

The use of machine learning for sensor signal interpretation in NDT reaches more than 25 years back [11]. The application of support vector machines and neural networks has a long history in the automatic evaluation of ultrasonic data and showed promising results [12]–[20]. However, finding the optimal setup for a particular purpose is difficult due to the wealth of available methods and the large number of hyper-parameters to be tuned.

Manuscript received December 24, 2020; revised February 23, 2021; accepted February 23, 2021. Date of publication March 1, 2021; date of current version April 5, 2021. This work was supported in part by the Ministry of Science, Research and Culture of Brandenburg through the Project Kognitive Materialdiagnostik under Grant 22-F241-03-FhG/007/001. The associate editor coordinating the review of this article and approving it for publication was Prof. Guiyun Tian. *(Corresponding author: Ivan Kraljevski.)*

Ivan Kraljevski, Frank Duckhorn, and Constanze Tschöpe are with Fraunhofer IKTS, Fraunhofer Institute for Ceramic Technologies and Systems, 01277 Dresden, Germany (e-mail: ivan.kraljevski@ikts.fraunhofer.de).

Matthias Wolff is with the Chair of Communications Engineering, Brandenburg University of Technology Cottbus–Senftenberg, 03046 Cottbus, Germany (e-mail: matthias.wolff@b-tu.de).

Digital Object Identifier 10.1109/JSEN.2021.3062941

Another issue is the amount of representative data required for reliable machine learning. Typically, there is an abundance of available data provided by intact specimens and little or none provided by the damaged ones.

This paper is based on preliminary work [21]–[23], in which we investigated traditional machine learning approaches, hidden Markov models (HMM) and support vector machines (SVM), for ultrasonic SHA of airplane materials.

Here, the severity of damage was predicted by different multi-/one-class classification and regression approaches.

All investigated methods performed reasonably well where the multi-class classification was particularly successful. However, the regression and one-class tasks were more difficult, but still practically feasible.

In this study, we applied state-of-the-art machine learning algorithms, namely, autoencoders and isolation forest, to the same data used in the aforementioned studies and we compared the results where applicable. The machine learning algorithms were employed in anomaly detection using exclusively sensor signals of an intact state.

Sensor signals obtained on artificial damage states were used in the detection of and damage severity assessment. The performance was evaluated and compared across different classifiers and feature analysis methods. We investigated the following anomaly detection methods: hidden Markov models (HMM), support vector machines (SVM), isolation forest (IF) and reconstruction autoencoders (AEC).

The paper is organized as follows: In Section II we describe the data collection, the datasets, and the feature analysis algorithms. Section III presents the anomaly detection methodology used for damage recognition as well as for severity assessment. Section IV outlines the datasets preparation, hyper-parameter optimization, a brief description of the ML algorithms, and decision fusion. Section V presents detailed analysis of the classification experiments.

II. Materials and Methods

A. Data Collection

We used two objects in our experiments, an aluminum plate that measured 1000 mm × 1000 mm × 2.5 mm (Fig. 1) and a carbon fiber reinforced plastic plate with dimensions of 860 mm (Fig. 2) [21], [22].

The data were collected in a controlled manner, starting with intact test objects and gradually adding damage. This procedure yielded plenty of data representing damaged states. This provides us with a much better possibility to study outlier detection than using data from a real use case where damaged states are rare.

1) Aluminum Plate (ALU): The aluminum (**ALU**) plate was equipped with $N = 8$ ultrasound transducers attached at the plate in a circular arrangement with a diameter of 570 mm (Fig. 1). During one measurement, the plate was excited by each of the eight transducers subsequently, while the remaining seven transducers were recording the arriving sound waves. Hence, a measurement comprises of $N \times (N - 1)$ or $8 \times 7 = 56$ signals from as many *signal paths*. A signal path is defined by the sending and receiving transducer and we simply designate it as a *sensor* (e.g. A2B1, D2B1, D2A2, etc.). As an excitation signal, we used a Ricker wavelet [24] with a center frequency of 250 kHz (**R250**). All measured signals have the same duration of 400 µs and were recorded with a sample rate of 6.25MS/s and 16-bit resolution. Table I shows the number of measurements per state and the total count of the signal recordings.

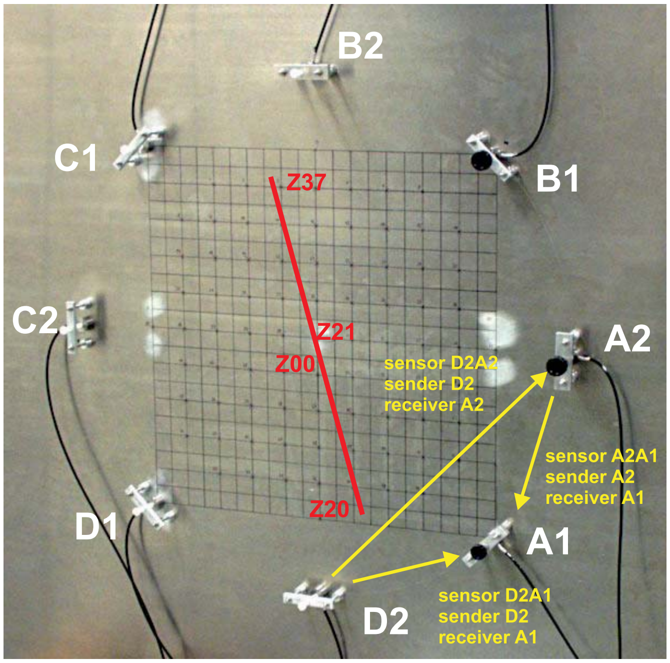


Fig. 1. Ultrasonic transducer configuration on the aluminum plate.

TABLE I
Aluminum Plate (ALU) Database

State	#Measurements	#Signals
Z00	2000	112 000
Z01	2000	112 000
Z02…Z05	4 × 100	4 × 5600
Z06	2000	112 000
Z07…Z37	31 × 100	31 × 5600
Sum	9500	532 000

The initial intact state of the aluminum plate is labeled as “Z00”. During the experiment, we introduced a fissure of increasing length in the center of the sensor arrangement. The resulting damage states are labeled as “Z01”…“Z37”, where the digits correspond to the fissure length in centimeters (see also [22, Fig. 10]).

2) Carbon Fiber Reinforced Plastic (CFRP): The **CFRP** plate was equipped with 12 ultrasound transducers in a 600×400mm grid arrangement (Fig. 2).

The employed measurement procedure was the same as for the aluminum plate, where for the CFRP plate a measurement is comprised of $12 \times 11 = 132$ signals (see also [22, Fig. 11]).

Similarly, the measured signals have an equal duration of 1046 µs. We used a sampling rate of 4.16MS/s and 16-bit resolution for the measurements and we tested three different excitation signals: Ricker wavelet [24] with center frequencies of 100 kHz (**R100**) and 350 kHz (**R350**), as well as a **sinc** function with a cutoff frequency of 600 kHz (**S600**).

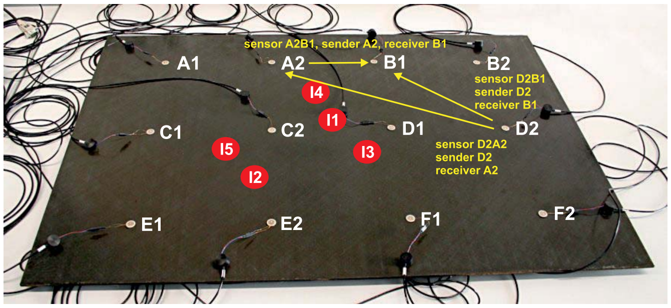


Fig. 2. Ultrasonic transducer configuration on CFRP plate.

TABLE II
ARTIFICIALLY INTRODUCED IMPACT DAMAGE

Impact	Energy (J)	Back	State	Damage
			Z00…Z02	none
I1	15	supported	Z03	I1
I2	25	supported	Z04	I1, I2
I3	45	supported	Z05	I1, I2, I3
I4	25	free	Z06	I1, I2,…, I4
I5	25	free	Z07	I1, I2,…, I5

TABLE III
CARBON FIBER REINFORCED PLASTIC PLATE (CFRP) DATABASE

Excitation	State	#Measurements	#Signals
R100	Z00, Z03…Z07	132 × 6 × 1000	792 000
R350	Z00…Z07	132 × 8 × 1000	1 056 000
S600	Z00, Z03…Z07	132 × 6 × 1000	792 000
Sum		132 × 20 000	2 640 000

In the beginning, we performed three series of measurements on the intact state labeled as "Z00" …"Z02". Then we introduced increasing damage by subsequently hitting the plate with a steel ball applying varying energy (15 J to 45 J) where the plate fixed on a flat surface for I1 …I3 (supported) and for I4 and I5 freely laid on a frame (free). After each hit a new damage state was introduced and the measurements performed (Table II).

This procedure resulted in five states of increasing damage severity for all excitation signals labeled as "Z03"…"Z07". The intact states "Z01" and "Z02" were recorded only with the R350 excitation signal. Table III summarizes the data collected from the CFRP plate.

B. Feature Analysis

We compared the classification of raw signals (**SIG**) and two feature extraction methods, "primary" (**PFA**) and "secondary" (**SFA**) features.

1) SIG: We used the raw recordings for classification without any pre-processing as features. Because all signals per test object have the same duration we obtain features with the same dimensions: 2500 (ALU) and 4352 (CFRP) signal samples.

2) PFA: Primary features (**PFA**) were obtained from the signals using short-time Fourier transform with a window length of 1024 signal samples and a continuation rate of 32 signal samples, followed by spectral sub-sampling by a factor 4 and subsequent low-pass filtering. The resulting feature dimensions are 47 × 24 for the ALU and 105 × 32 for CFRP plate.

3) SFA: Secondary features (**SFA**) were computed from the primary features by using vector standardization to zero mean and unit variance. Additionally, principal component analysis (PCA) was employed for reduction to 16 spectral dimensions. Therefore, the dimension of the SFA features was 47 × 16 for ALU and 105 × 16 for CFRP.

III. DEFECT DETECTION AND ASSESSMENT

A. Anomaly Detection

In a typical application, we have plenty of sensor measurements indicating intact state and much less of those of structural defects. Therefore the classifiers are trained to detect anomalies by using only the signals of the intact (normal) state in a one-class classification scenario.

We can define a data set as:

$$\mathcal{D} = \left\{(x^{(1)}, y^{(1)}), \dots, (x^{(N)}, y^{(N)})\right\} \tag{1}$$

consisting of $N = |\mathcal{D}|$ feature-label pairs $(x, y) \in \mathcal{D}$ called *samples*. The ground truth labels are $y \in \{c_0, c_1\}$, where c_0 denotes an anomalous and c_1 a normal state. The data set is partitioned according to the ground truth labels:

$$\begin{aligned}\mathcal{D}_0 &= \left\{(x, y) \in \mathcal{D} : y = c_0\right\} \text{ – labeled as } c_0,\\ \mathcal{D}_1 &= \left\{(x, y) \in \mathcal{D} : y = c_1\right\} \text{ – labeled as } c_1.\end{aligned} \tag{2}$$

where $\mathcal{D} = \mathcal{D}_0 \cup \mathcal{D}_1$ and $\mathcal{D}_0 \cap \mathcal{D}_1 = \emptyset$.

The classifiers compute scores for each sample (likelihoods, decision functions, etc.) which were transformed into a pseudo posterior probability $P(c_1|x) \in [0, 1]$ of the sample x belonging to the "normal" state's class c_1.

A sample is considered anomalous—class c_0—if this pseudo-probability falls below an estimated threshold. Details on threshold computation and the calculation of pseudo-probabilities, which differs between the classifiers, are given below.

Since our dataset is highly imbalanced, evaluation with the basic accuracy rate would yield biased results [25]. A more appropriate performance indicator is the posterior balanced accuracy rate [26], [27].

The accuracy rate for each class i (the recall) is defined as:

$$A_i = \frac{C_i}{N_i}, \quad i \in \{0, 1\}, \tag{3}$$

where $N_i = |\mathcal{D}_i|$ is the total number of samples and C_i is the number of correctly classified samples for class c_i. According to [26], the posterior distribution of the class accuracy rate A_i can be expressed as a Beta distribution with parameters α_i and β_i:

$$(A_i|C_i) \sim \text{Beta}(\alpha_i, \beta_i), \quad i \in \{0, 1\}, \tag{4}$$

where $\alpha_i = 1 + C_i$ and $\beta_i = 1 + N_i - C_i$, and the probability density x of correctly classifying unseen samples is:

$$P_{A_i}(x; \alpha_i, \beta_i) = \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} x^{\alpha_i - 1}(1-x)^{\beta_i - 1}, \ i \in \{0, 1\}. \tag{5}$$

The balanced accuracy rate is estimated as the average of the individual recalls $\frac{1}{2}(A_0 + A_1)$, whose probability density can be estimated by convolution:

$$P_B(x; \alpha_0, \beta_0, \alpha_1, \beta_1) = \int_0^1 P_{A_0}\big(2(x-y); \alpha_0, \beta_0\big) \cdot P_{A_1}\big(2y; \alpha_1, \beta_1\big)\, dy. \tag{6}$$

$$BAR = E|X| = \int_0^1 x \cdot P_B(x)\, dx \tag{7}$$

From here, the balanced accuracy is the expected value (7) of the probability density (5). We compute the 95% Clopper-Pearson confidence interval (CI) according to [28]. In this case, after balancing, the random guess baseline for the accuracy rate is 50% which represents the inverse number of classes.

We also calculate the equal error rate (EER) with the 95% CI and the corresponding threshold. Such threshold obtained on a development set can be used for the classifier calibration providing more reliable anomaly detection.

Additionally, when an ideal detection is achieved ($EER = 0$), the classification (safety) margin (CM) is used as a metric of the robustness of the classifiers:

$$CM = \frac{\min_{x\in\mathcal{D}_1} P(c_1|x) - \max_{x\in\mathcal{D}_0} P(c_1|x)}{\text{mean}_{x\in\mathcal{D}_1} P(c_1|x) - \text{mean}_{x\in\mathcal{D}_0} P(c_1|x)}, \tag{8}$$

where $\text{mean}_{x\in\mathcal{D}_1} P(c_1|x) > \text{mean}_{x\in\mathcal{D}_0} P(c_1|x)$. The safety margin represents the smallest relative distance of any two classes and it is always between zero and one.

B. Severity Assessment

The damage severity assessment is performed using the knowledge about the true labels (y) of the anomalous states as described in Tables I and III. The objective is to investigate how good the models trained only on signals of intact state can estimate the damage severity for a specific state (like the length of the fissure in the ALU or the damage level in the CFRP database). The damage severity assessment makes sense only if we have ideal anomaly detection and an existing development set which is needed for modeling and calibration.

One possible approach is to perform linear regression modeling on the fused pseudo-probability scores across the ground truth labels (see details below), comparing the mean square error (MSE) and the coefficient of determination (R-square or R2) across different classifiers. These metrics will give an insight about the linearity of the scores of one state label, but it will not give the information about the score dispersion inside and outside the label states.

Another approach which will complement the severity assessment is to consider the scores with their true labels as the outcome of a cluster modeling, and hence to use appropriate performance metrics like silhouette coefficient (SC). SC is the function of the mean distance between a score and all other scores of the same label (tightness) and the mean distance between a score and all other scores in the next nearest cluster (separation) [29]. Such kind of metrics provide qualitative comparison of classifiers regarding their ability for damage severity assessment.

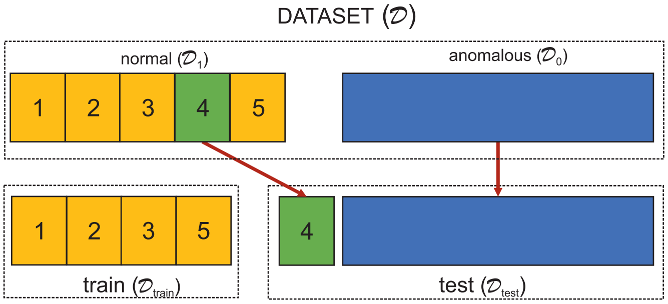

DATASET ($\mathcal{D}$)
normal ($\mathcal{D}_1$)
anomalous ($\mathcal{D}_0$)
1
2
3
4
5
1
2
3
5
4
train ($\mathcal{D}_{train}$)
test ($\mathcal{D}_{test}$)

Fig. 3. Partition into training and test set for fold 4.

IV. Experimental Setup

We investigate the performance of the classifiers on both test objects (ALU and CFRP) and with all excitation signals, giving in total four different datasets (see Table I and III).

As already mentioned, in a typical application little or no data of damage states are available. It is impossible to cover a large variety of possible defect states. Therefore, the training set was composed exclusively of signals representing the intact state of the objects.

A. Dataset Preparation

We train one model for every signal path (or sensor) with stratified 5-fold cross-validation. Reproducible training runs were ensured by fixing the random generator seed. The fold divisions for the training were kept the same across a signal path (sensor) for all classifiers and feature sets. We define a data set partition for each fold into training $\mathcal{D}_{train}$ and test set $\mathcal{D}_{test}$, where $\mathcal{D} = \mathcal{D}_{train} \cup \mathcal{D}_{test}$ and $\mathcal{D}_{train} \cap \mathcal{D}_{test} = \emptyset$. The train dataset test fold (Fig. 3) was combined with the dataset portion of the "anomalous" signals, "Z01"…"Z37" for the ALU and "Z03"…"Z07" for the CFRP datasets. This allows us to assess the performance in terms of false positives and false negatives for both classes.

B. Hyper-Parameters Selection

In all datasets (one for ALU and three for CFRP), and for all classifier-feature combinations, selection of hyper-parameters was performed by random search over a range of feasible parameter values on a small development set.

For classifiers with few parameters (one-class SVM, isolation forest and HMMs) the procedure can be expected to find close to optimal values. For the autoencoders with many parameters, the random search discovered values which might not be optimal, but the performance metrics were not much different. However, due to the relatively small dimensionality of the primary and secondary features, as well as high separability of the classes, the optimization is quite robust and the discovered parameters and the network architectures were relatively simple.

C. Machine Learning Approaches

We employed classification methods which are suitable for anomaly detection. For all of them, a single observation is classified by the corresponding model into two classes labeled

c_0 = "0" for anomaly and c_1 = "1" for normal states. The prediction was made according to the estimated threshold or to a given decision function of the classifier.

1) Autoencoder: Autoencoders (AEC) are deep neural networks whose architecture consists of two parts, an encoder which transforms the input to a compressed representation and the decoder which approximately reconstructs the input data, as well as possible according the learned representation (under-complete autoencoder) [30]. The autoencoder contains a bottleneck hidden layer, which forces the network to learn the salient features of the data and prevents the inputs from simply being passed to the output. To train and evaluate the autoencoder, we used Keras [31] with Tensorflow [32] as its back-end. The autoencoder was trained with the objective to reconstruct the input samples as well as possible, by minimizing the mean-squared error (MSE) as the loss function.

If an anomalous sample is tested, the trained autoencoder will fail to reconstruct it properly, consequently producing a higher than usual MSE. The MSE is considered to be a deviation score from the "normal state" and $MSE(x) \in [0, \infty)$. Pseudo-probabilities are estimated by applying a sigmoid function of the anomaly scores:

$$P(c_1|x) = \frac{1}{1 + \mathrm{e}^{-score(x)}}, \quad x \in \mathcal{D}_{test}. \tag{9}$$

where the anomaly score is defined:

$$score(x) = MSE(x). \tag{10}$$

We performed a random search over the initial network architecture—consisting of 3 hidden fully connected layers—varying the units in the layers while always setting the number of units in the bottleneck layer smaller than the preceding and the following layers. The batch size, the dropout rate and the optimization function were also investigated and optimized. To avoid over-fitting on such a small network on the training fold, we used early stopping with the criteria of 25 epochs without an improvement in the MSE score on the validation set. In this case, the validation set was randomly selected for each epoch as 20 percent of the training set.

2) Hidden Markov Models: Hidden Markov models (HMM) are statistical models for time-varying sequences. Apart from other areas of applications like speech and handwriting recognition, they were also successfully applied in non-destructive testing and technical signals evaluation [22].

The HMMs were generated with the dLabPro software [33], [34] using 3 emitting states and full covariance matrices. We used negative log-likelihood transformed into a pseudo-probability describing how well a model fits an observation. The pseudo-probability scores were calculated as antilogarithm of the neg-log. likelihoods $NLL(x)$ of test samples x:

$$P(c_1|x) = 10^{-\dfrac{NLL(x)}{\gamma}}, \quad x \in \mathcal{D}_{test}. \tag{11}$$

To avoid having a case where resulting pseudo-probabilities are near zero or near one, we use γ parameter estimated as the mean of neg-log. values of the N samples in the test set:

$$\gamma = \frac{1}{|\mathcal{D}_{test}|} \sum_{y \in \mathcal{D}_{test}} NLL(y). \tag{12}$$

3) One-Class SVM: Support vector machines [35], although developed for binary classification, are also used for anomaly detection as their natural extension for unlabeled data [36], [37]. We used the LIBSVM library [38] and the python scikit-learn [39] interface. Important hyper-parameters of the classifier are the kernel type and the ν which represents the upper bound of the fraction of training errors and the lower bound of the fraction of support vectors. The optimal values related to prediction accuracy were discovered with random search optimization. We used the radial basis function (RBF) kernel because of its good general performance in estimation of the support of a high-dimensional distribution.

The anomaly scores are obtained by the decision function which represents the signed distance to the separating hyperplane. When the score has a value greater than or equal to zero, the input sample is considered to be normal, otherwise, it is considered to be an anomaly. From here the pseudo-probability scores were estimated by the signed distances and the sigmoid function (9).

4) Isolation Forest: Isolation forest (IF) or iForest [40] is a method that, instead of profiling the normal samples, isolates the anomalies.

The measure of normality (decision function) is averaged over a forest of random trees and expressed as the number of recursive splits necessary to isolate a sample. The anomaly score for the samples is estimated as described in [41] where the lower values represent more abnormal points. The scores are transformed into pseudo-probabilities using (9).

The method performs well for large datasets, high dimensional features which are usually highly redundant, and for training sets which do not contain any anomalies. As in the case of one-class SVMs we used the python scikit-learn implementation of iForest.

D. State Label Predictions

Different classification models have their characteristic way of providing pseudo posterior probabilities of a sample being "normal", as described in Section IV-C.

For each signal path (56 for ALU and 132 for CFRP) a separate model was trained which provides a pseudo-probability score for an input observation that originates from an intact object state. To predict the state labels, it is necessary to compare the pseudo-probability score against the model-specific threshold.

The threshold θ was determined by the interquartile range rule (IQR) applied to the pseudo-probability scores obtained from observations of the intact state (i.e., the training set). We prefer this approach because it is a non-parametric method that does not assume the normal distribution and because it is more robust than the Z-score method.

Thus, we set the threshold θ to:

$$\theta = Q1 - 1.5 \cdot IQR, \text{ where } IQR = Q3 - Q1. \tag{13}$$

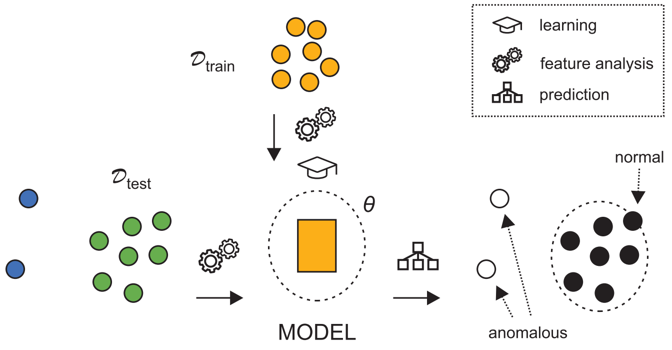


Fig. 4. Schematic view of model training and label prediction.

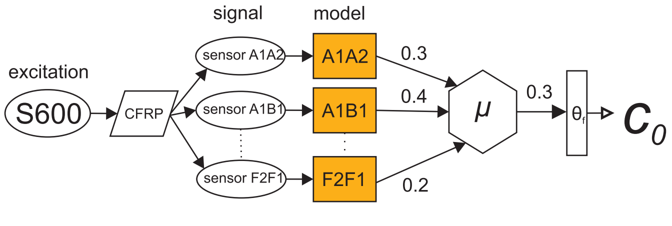


Fig. 5. Label prediction of an ensemble of estimators (CFRP-S600).

Q1 is the median value of the first and Q3 is the median values of the second half of the rank-ordered pseudo-probabilities.

The predicted label $\hat{y}$ for sample x is decided by the threshold θ:

$$\hat{y} = \begin{cases} c_0 \text{ iff } P(c_1|x) < \theta \text{ and} \\ c_1 \text{ otherwise.} \end{cases} \tag{14}$$

E. Sensor Fusion

The fusion of the sensors was performed on the decision level (late fusion), since fusion on the signal or feature level (early fusion) would create datasets with a huge feature dimensionality, which for the given sample count would render model training difficult or infeasible.

We combine the separate sensor models into an ensemble of classifiers for label prediction (Fig. 5). The resulting fused pseudo-probability is calculated as a weighted average of the individual pseudo-probabilities provided by all sensor models (soft voting). Here we assume that all the sensors are of equal importance.

For a set of sensors $S = \{\text{A1A2}, \text{A1B2}, \text{A1C1}, \ldots\}$ the fused pseudo-probability for sample x is given by:

$$\overline{P}(c_1|x) = \frac{1}{|S|} \sum_{s \in S} P_s(c_1|x), \tag{15}$$

where $P_s(c_1|x)$ denotes the pseudo-probability computed for sensor s.

Since we have a one-class classification scenario, the fused pseudo-probability is compared to a threshold θ_f for label prediction (14).

The threshold is calculated on fused pseudo-probabilities, as with individual sensor models (13).

From here, we can compute other metrics (defined in Section III) like EER, the CM, the MSE, the R2 and, SC for damage severity assessment.

TABLE IV
BALANCED ACCURACY RATE (95% CI) (NO RECALIBRATION)

Classifier-Feature		EXPERIMENT			
		ALU-R250	CFRP-R100	CFRP-R350	CFRP-S600
AEC	PFA	$98.251^{+0.374}_{-0.426}$	$98.752^{+0.423}_{-0.527}$	$99.334^{+0.191}_{-0.209}$	$99.152^{+0.373}_{-0.427}$
AEC	SFA	$98.976^{+0.299}_{-0.351}$	$\mathbf{99.551^{+0.224}_{-0.326}}$	$96.572^{+0.253}_{-0.247}$	$98.353^{+0.522}_{-0.578}$
AEC	SIG	$96.702^{+0.523}_{-0.577}$	$98.154^{+0.521}_{-0.629}$	$98.251^{+0.324}_{-0.326}$	$98.453^{+0.472}_{-0.578}$
HMM	PFA	$98.976^{+0.299}_{-0.351}$	$\mathbf{99.551^{+0.224}_{-0.326}}$	$99.947^{+0.028}_{-0.072}$	$99.102^{+0.373}_{-0.477}$
HMM	SFA	$99.200^{+0.275}_{-0.275}$	$98.949^{+0.226}_{-0.324}$	$88.244^{+0.281}_{-0.319}$	$\mathbf{99.751^{+0.174}_{-0.226}}$*
IF	PFA	$97.376^{+0.449}_{-0.501}$	$\mathbf{99.451^{+0.274}_{-0.376}}$	$\mathbf{99.997^{+0.003}_{-0.072}}$*	$99.501^{+0.274}_{-0.376}$
IF	SFA	$99.200^{+0.275}_{-0.275}$	$99.202^{+0.273}_{-0.377}$	$88.450^{+0.325}_{-0.325}$	$99.501^{+0.274}_{-0.376}$
IF	SIG	$97.455^{+0.420}_{-0.480}$	$98.159^{+0.316}_{-0.434}$	$88.557^{+0.318}_{-0.332}$	$96.404^{+0.321}_{-0.429}$
SVM	PFA	$99.325^{+0.250}_{-0.300}$	$99.002^{+0.373}_{-0.477}$	$99.500^{+0.175}_{-0.175}$	$99.102^{+0.373}_{-0.477}$
SVM	SFA	$\mathbf{99.675^{+0.150}_{-0.200}}$*	$\mathbf{99.351^{+0.274}_{-0.376}}$	$91.615^{+0.310}_{-0.340}$	$99.451^{+0.274}_{-0.376}$
SVM	SIG	$95.975^{+0.250}_{-0.250}$	$98.114^{+0.411}_{-0.489}$	$91.881^{+0.344}_{-0.356}$	$98.885^{+0.290}_{-0.410}$

*significantly ($p < 0.05$) different prediction performance compared across classifier-feature combinations in same experiment (column)

F. Recalibration

When signal samples of either naturally or artificially damaged objects become available, it will be possible to recalibrate the classifiers and maximize the detection performance.

Labeled signals of both states (intact and damaged) were evaluated the same way as described in Section IV-D. Then the resulting scores were used to estimate the EER with its 95% CI according to the procedure described in Section III-A. This metric provides the optimal pseudo-probability threshold θ_f for label predictions and is an indicator for the anomaly detection power of the trained models.

V. RESULTS AND DISCUSSION

A. Anomaly Detection Results

Table IV presents the results of damage detection in terms of BAR with the 95% confidence intervals without recalibration for all the datasets across features and classifiers. The labels, i.e. normal (c_1) and anomalous (c_0), were predicted as described in Section IV-E.

Most of the classifier-feature combinations performed very well. The combination with the highest expected BAR is marked for each dataset.

McNemar's test [42] showed that the top feature-classifier combinations significantly outperform the others ($p < 0.05$), except for the CFRP-R100 experiment where we found four top combinations significantly outperforming all others, but not significantly different from each other.

Fig. 6 presents the results for the equal error rate (EER) after recalibration as described in Section IV-F. Hence, the performances cannot be directly compared with those presented in Table IV since the prediction thresholds are differently calculated.

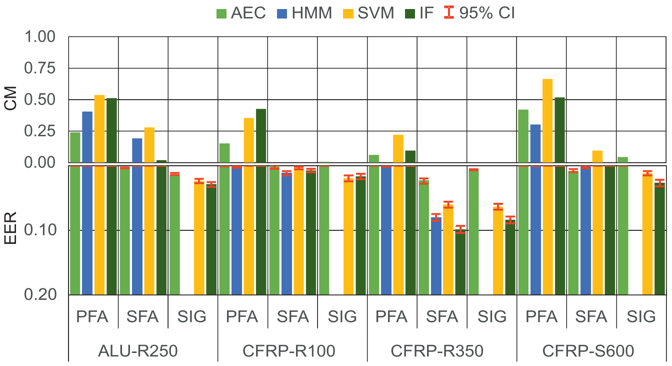


Fig. 6. Equal error rates (lower) and the safety margins (upper plot) (with recalibration).

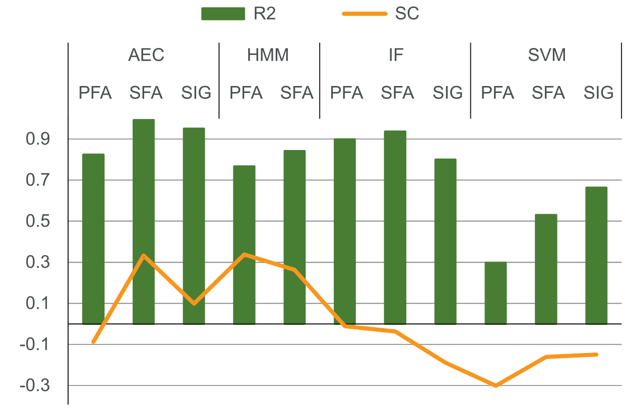


Fig. 7. Coefficient of determination and the silhouette coefficient on the ALU dataset.

The error rates are plotted in the lower part of the figure (the higher the bar the lower the error rate). The upper part of the figures presents the safety margins (CM) according to the (8). They are calculated only when the error rate is zero and they indicate the robustness of the classifier-feature combinations, the higher the bar the better the result.

In the case of the aluminum plate with 250 kHz Ricker excitation signal for the PFA features all classifiers achieved EER of zero and more than 20% safety margin. For the SFA features, the autoencoder approach failed to provide error-free anomaly detection, the other achieved margins in the range of 2−29%. For the raw signals, all of the tested methods have an EER not less than 3% which is still a solid anomaly detection performance (note: HMM classifiers with raw signals are infeasible).

In the case of the CFRP plate, all of the methods trained with PFA features of the S600 dataset achieved zero EER with comfortable margins of more than 30%.

For the same features, only HMMs with Ricker excitation signals R100 and R350 failed to reach the zero EER, however, the margins for the other were much lower than those achieved with S600.

Classifiers combined with SFA and SIG did not reach the zero EER except the AEC-SIG with R100 ($CM = 0.6\%$) and with S600 the SVM-SFA ($CM = 9.7\%$) and AEC-SIG ($CM = 4.4\%$).

It is notable that only the autoencoder approach was able to reliably detect anomalies using only the raw signals without additional feature analysis.

B. Severity Assessment Results

In regard to damage assessment, we have two different types of introduced defects, as described in Section II-A. In the ALU-R250 dataset the damage is a fissure of increasing length from 1−37 centimeters. The damage severity scores could be modeled by linear regression on the pseudo-probabilities and the quality of the model evaluated by the mean-squared error (MSE), coefficient of determination R2, and the silhouette coefficient (SC).

Fig. 7 shows that the AEC-SFE classifier-feature combination yielded the best linear regression fit with low dispersion and the lowest MSE of 0.886 (Fig. 8).

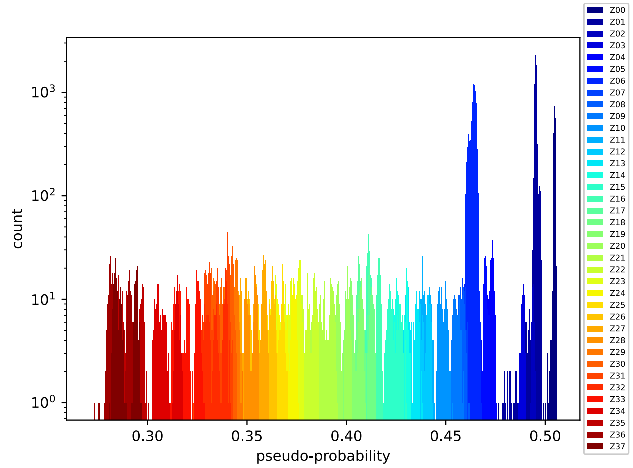


Fig. 8. Pseudo-probabilities of AEC-SFA combination (ALU-R250).

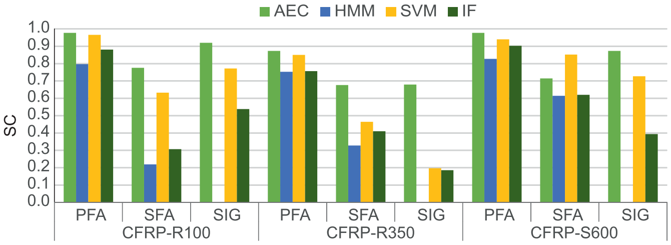


Fig. 9. Silhouette coefficients on the CFRP datasets.

For the CFRP dataset the damage was introduced by impacts with different energies on the plate (see Section II-A.2), which cumulative add to de damage in a non-linear fashion. Hence, linear regression would not properly model the damage severity. In this case, it is more informative to present the silhouette coefficient that estimates the quality of the pseudo-probability density clusters in terms of cluster tightness and their separation. In both cases, in practical applications, only a few samples of defect signals are necessary to reliably recalibrate the mapping of the pseudo-probability densities and to provide precise quantitative estimation of the damage severity scores.

Fig. 9 presents the silhouette coefficients of the anomaly score distributions on the CFRP database. The results correspond to those for EER and CM, the PFA features provided the best damage severity scores clustering across all classifiers

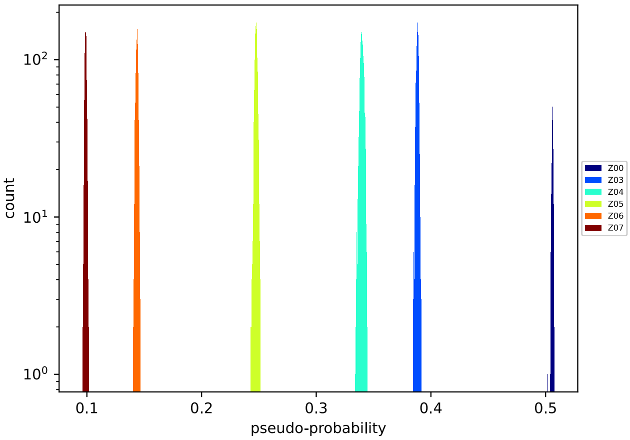


Fig. 10. Pseudo-probabilities of AEC-PFA combination (CFRP-S600).

and excitation signals (Fig. 10). It has to be emphasized that autoencoder models provided much better clustering scores than the other for the SFA and SIG features and all types of excitation signals.

VI. Conclusion

This study builds upon previous work and presents an extended robust methodology for anomaly detection and assessment. We presented a novel approach for anomalous dataset preparation and damage severity assessment utilizing exclusively information of an intact object state.

The experiment results show that we were able to train not only accurate classifiers but we also achieved a considerable safety margin. We employed various machine learning algorithms for anomaly detection with different feature analyses on ultrasonic sensor measurements. We recorded the responses of various excitation signals in a sensor network on two structures of different materials commonly used in various industries, such as aerospace and automotive.

Finally, we demonstrated that reliable anomaly detection is practically feasible, even if only data of the intact object state are available. When recalibration data becomes available either during operation or by expert knowledge, then accurate quantitative analysis of damage severity is possible as well.

References

[1] S. N. Atluri, S. G. Sampath, and P. Tong, *Structural Integrity Aging Airplanes*. Berlin, Germany: Springer, 2012. [Online]. Available: https://www.springer.com/gp/book/9783642843662

[2] M. Namkung, B. Wincheski, and N. Padmapriya, "NDT in the aircraft and space industries," in *Reference Module in Materials Science and Materials Engineering*. Amsterdam, The Netherlands: Elsevier, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780128035818019408

[3] H. Towsyfyan, A. Biguri, R. Boardman, and T. Blumensath, "Successes and challenges in non-destructive testing of aircraft composite structures," *Chin. J. Aeronaut.*, vol. 33, no. 3, pp. 771–791, Mar. 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1000936119303474

[4] H. Tretout, "Review of advance ultrasonic techniques for aerospace structures," in *Proc. Eur. Conf. Non-Destructive Test. (ECNDT)*, vol. 7, 1998, pp. 318–325.

[5] W. Staszewski, C. Boller, and G. R. Tomlinson, *Health Monitoring of Aerospace Structures: Smart Sensor Technologies and Signal Processing*. Hoboken, NJ, USA: Wiley, 2004.

[6] G. Riegert, K. Pfleiderer, H. Gerhard, I. Solodov, and G. Busse, "Modern methods of NDT for inspection of aerospace structures," in *Proc. Eur. Conf. Non-Destructive Test. (ECNDT)*, vol. 9, 2006, pp. 1–11.

[7] L. W. Schmerr, *Fundamentals of Ultrasonic Nondestructive Evaluation*. Cham, Switzerland: Springer, 2016. [Online]. Available: https://www.springer.com/gp/book/9783319304618

[8] R. D. Finlayson, M. Friesel, M. Carlos, P. Cole, and J. Lenain, "Health monitoring of aerospace structures with acoustic emission and acousto-ultrasonics," *Insight-Wigston Then Northampton*, vol. 43, no. 3, pp. 155–158, 2001.

[9] B. W. Drinkwater and P. D. Wilcox, "Ultrasonic arrays for non-destructive evaluation: A review," *NDT E Int.*, vol. 39, no. 7, pp. 525–541, Oct. 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0963869506000272

[10] I. Virkkunen, T. Koskinen, O. Jessen-Juhler, and J. Rinta-Aho, "Augmented ultrasonic data for machine learning," 2019, *arXiv:1903.11399*. [Online]. Available: http://arxiv.org/abs/1903.11399

[11] J. C. Royer, A. Merle, and C. de Sainte Marie, "An application of machine learning to the problem of parameter setting in non-destructive testing," in *Proc. 3rd Int. Conf. Ind. Eng. Appl. Artif. Intell. Expert Syst. (IEA/AIE)*, 1990, pp. 972–980.

[12] S. J. Farley, J. F. Durodola, N. A. Fellows, and L. H. Hernández-Gómez, "High resolution non-destructive evaluation of defects using artificial neural networks and wavelets," *NDT E Int.*, vol. 52, pp. 69–75, Nov. 2012.

[13] M. Bilgehan and P. Turgut, "The use of neural networks in concrete compressive strength estimation," *Comput. concrete*, vol. 7, no. 3, pp. 271–283, Jun. 2010.

[14] T. D'Orazio, M. Leo, A. Distante, C. Guaragnella, V. Pianese, and G. Cavaccini, "Automatic ultrasonic inspection for internal defect detection in composite materials," *NDT E Int.*, vol. 41, no. 2, pp. 145–154, Mar. 2008.

[15] S. Sambath, P. Nagaraj, and N. Selvakumar, "Automatic defect classification in ultrasonic NDT using artificial intelligence," *J. Nondestruct. Eval.*, vol. 30, no. 1, pp. 20–28, Mar. 2011.

[16] S. Seyedtabaii, "Performance evaluation of neural network based pulse-echo weld defect classifiers," *Meas. Sci. Rev.*, vol. 12, no. 5, pp. 168–174, Jan. 2012.

[17] Y. Ying *et al.*, "Toward data-driven structural health monitoring: Application of machine learning and signal processing to damage detection," *J. Comput. Civil Eng.*, vol. 27, no. 6, pp. 667–680, Nov. 2013.

[18] E. Simas Filho, M. M. Silva, Jr., P. C. Farias, M. C. Albuquerque, I. C. Silva, and C. T. and Farias, "Flexible decision support system for ultrasound evaluation of fiber–metal laminates implemented in a DSP," *NDT E Int.*, vol. 79, pp. 38–45, Apr. 2016.

[19] M. Meng, Y. J. Chua, E. Wouterson, and C. P. K. Ong, "Ultrasonic signal classification and imaging system for composite materials via deep convolutional neural networks," *Neurocomputing*, vol. 257, pp. 128–135, Sep. 2017.

[20] O. Mac Aodha *et al.*, "Bat detective—Deep learning tools for bat acoustic signal detection," *PLoS Comput. Biol.*, vol. 14, no. 3, 2018, Art. no. e1005995.

[21] C. Tschope, E. Schulze, H. Neunubel, M. Wolff, R. Schubert, and R. Hoffmann, "Experiments in acoustic structural health monitoring of airplane parts," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 2037–2040.

[22] C. Tschope and M. Wolff, "Statistical classifiers for structural health monitoring," *IEEE Sensors J.*, vol. 9, no. 11, pp. 1567–1576, Nov. 2009.

[23] C. Tschoepe, F. Duckhorn, and M. Wolff, "Comparison of machine learning algorithms for NDT in aerospace," in *Proc. 10th Int. Symp. NDT Aerosp.*, Dresden, Germany, Oct. 2018, p. 17.

[24] N. Ricker, "The form and laws of propagation of seismic wavelets," *Geophysics*, vol. 18, no. 1, pp. 10–40, Jan. 1953.

[25] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognit.*, vol. 91, pp. 216–231, Jul. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320319300950

[26] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 3121–3124.

[27] A. Tran, C. S. Ong, and C. Wolf, "Combining active learning suggestions," *PeerJ Comput. Sci.*, vol. 4, p. e157, Jul. 2018.

[28] C. J. Clopper and E. S. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, vol. 26, no. 4, pp. 404–413, 1934.

[29] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0377042787901257

[30] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[31] F. Chollet *et al.* (2015). *Keras*. [Online]. Available: https://github.com/fchollet/keras

[32] M. Martín *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Operating Syst. Design Implement. (OSDI)*, 2016, pp. 265–283.

[33] R. Hoffmann, M. Eichner, and M. Wolff, "Analysis of verbal and nonverbal acoustic signals with the Dresden UASR system," in *Int. Workshop Verbal Nonverbal Commun. Behaviours. COST Action* (Lecture Notes in Computer Science), A. Esposito, M. Faundez-Zanuy, E. Keller, and M. Marinaro, Eds., vol. 4775. Vietri sul Mare, Italy: Springer-Verlag, Mar. 2007, pp. 200–218. [Online]. Available: http://www.springerlink.com/content/e252731142457687/

[34] M. Wolff. (2014). *dLabPro: A Signal Processing and Acoustic Pattern Recognition Toolbox*. [Online]. Available: https://github.com/matthias-wolff/dLabPro

[35] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[36] B. Schölkopf, R. C. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, "Support vector method for novelty detection," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 12, 1999, pp. 582–588.

[37] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, Jul. 2001.

[38] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, p. 27, 2011.

[39] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[40] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 413–422.

[41] F. T. Liu, K. M. Ting, and Z. Zhou, "Isolation-based anomaly detection," *ACM Trans. Knowl. Discovery from Data*, vol. 6, no. 1, pp. 1–39, Mar. 2012.

[42] S. L. Salzberg, "On comparing classifiers: Pitfalls to avoid and a recommended approach," *Data Mining Knowl. Discovery*, vol. 1, no. 3, pp. 317–328, 1997.

Ivan Kraljevski was born in Emmen, The Netherlands, in 1974. He received the B.Sc. degree in electronics and telecommunications, the M.Sc. degree in computer technology and informatics, and the Ph.D. degree from the Faculty of Electrical Engineering and Information Technology, University "St. Cyril and Methodius," Skopje, Macedonia, in 1997, 2000, and 2004, respectively. From October 2011 to September 2013, he worked as a Researcher of Speech Technology with IAS, TU-Dresden. His current position is Research Associate with the Fraunhofer Institute for Ceramic Technologies and Systems (IKTS), Dresden, Germany. His scientific and professional interests include speech and audio signal processing, speech recognition and synthesis, speaker identification, noise-robust speech recognition, pattern recognition, and artificial neural networks.

Frank Duckhorn was born in Erlabrunn, Germany, in 1981. He received the Diploma degree in information system technology and the Ph.D. degree from the Dresden University of Technology, Germany, in 2007 and 2014, respectively. From 2007 to 2014, he worked as a Researcher of Speech Technology and Pattern Recognition with the Institute of Acoustic and Speech Communication, Dresden University of Technology. Since 2014, he works in the field of machine learning and data analysis for technical acoustic signals with the Fraunhofer Institute for Ceramic Technologies and Systems (IKTS), Dresden, Germany. He is interested in signal processing, acoustic signal, pattern recognition, machine learning, and hardware optimized algorithms.

Constanze Tschöpe received the Dipl.-Inf. degree in computer science from the Technische Hochschule Leipzig, Leipzig, Germany, in 1996, and the Ph.D. degree from the Faculty of Electrical Engineering and Information Technology, Dresden University of Technology (Technische Universität Dresden), Germany, in 2011. Since 1996, she has been a Research Associate with the Fraunhofer Institute for Ceramic Technologies and Systems (IKTS) (formerly: IZFP-EADQ und IZFP-D), Dresden, Germany. Since 2018, she has been the Head of the Machine Learning and Data Analysis Group and the Head of the Cognitive Material Diagnostics Project Group with the Brandenburg University of Technology, Cottbus. Her research interests include artificial intelligence, machine learning, acoustic pattern recognition, intelligent signal processing, and quality assessment of technical processes.

Matthias Wolff was born in Görlitz, Germany. He received the Dipl.-Ing. (M.Sc.) and Dr.-Ing. (Ph.D.) degrees in electrical engineering and information technology and the Habilitation degree in systems theory from TU Dresden in 1997, 2004, and 2011, respectively. Since 2011, he has been working as a Full Professor of Communications Engineering with the Brandenburg University of Technology Cottbus-Senftenberg, Germany. He is lecturing on systems theory, communications engineering, speech and language technology, and cognitive systems. His scientific interests are mainly in text and semantics processing, behavior control of cognitive machines, and quantum-inspired AI methods.