Acoustic Structural Integrity Assessment of Ceramics using Supervised Machine Learning and Uncertainty-Based Rejection*

Maria Lua Nunes†
Associação Fraunhofer Portugal Research
Porto, Portugal

Marília Barandas
Associação Fraunhofer Portugal Research
Porto, Portugal
FCT NOVA
Laboratory for Instrumentation
Caparica, Portugal

Hugo Gamboa
Associação Fraunhofer Portugal Research
Porto, Portugal
FCT NOVA
Laboratory for Instrumentation
Caparica, Portugal

Filipe Soares
Associação Fraunhofer Portugal Research
Porto, Portugal

ABSTRACT

Industry and Quality 4.0 pose the opportunity to integrate artificial intelligence-based technology into the quality management of products/services. Particularly, quality control procedures of tableware ceramics require a demanding and faulty human manual (visual and acoustic) inspection. In this paper, we propose an uncertainty-aware automated acoustic inspection using a supervised machine learning model based on a set of novel acoustic features to classify ceramic plates, as cracked and uncracked. We conducted experiments on a dataset of 31 ceramic plates (16 cracked and 15 uncracked), collected in the laboratory. Data quality check and augmentation strategies were also performed, resulting in 2900 samples. The main contributions of this paper are: 1) description of 192 features selected for the acoustic inspection of ceramic plates; 2) comparison of model calibration results regarding three different classifiers; 3) study of different sources of uncertainty for classification with rejection option, through uncertainty quantification measures, and the effect of feature selection on it. We performed two experiments that differ in the usage of a supervised feature selection method. We split the augmented dataset into train/test sets in a proportion of 90/10. The calibrated SVM was selected as the best classifier based on model calibration and cross-validation results and was used in the prediction on the test set. The uncertainty-based rejection improved the train and test sets' classification results. In the experiment with feature selection, the classification performance remained high, while the uncertainty about the predictions and the percentage of rejected samples decreased.

*This work was financially supported by the project Visual and Acoustics Inspection of Ceramics (VAICeramics), co-funded by Portugal 2020 framed under the Operational Programme for Competitiveness and Internationalization (COMPETE 2020), by Agência Nacional de Inovação and European Regional Development Fund under Grant POCI-01-0247-FEDER-069987.

†corresponding author: maria.nunes@fraunhofer.pt

Keywords

Ceramics; Quality Management; Defect Inspection; Time series; Artificial Intelligence

1. INTRODUCTION

Quality management of products or services within manufacturing processes guarantees products' reliability [7]. The current fourth Industrial Revolution is ruled by cyber-physical systems, which integrate the physical world with the advancements of the information era for forthcoming industrial procedures [13]. Industry 4.0 technologies, e.g. embedded devices, wireless communications, and methods of artificial intelligence, are leveraging quality management operations resulting in Quality 4.0, which is characterized by digitization [10] and Zero Defect Manufacturing (i.e. the set of methodologies and strategies for the elimination of defective components during production). Zero Defect Manufacturing reduces the operational costs associated with defective components, thus it is very appealing to industries [17].

Ceramic materials are a class of inorganic materials made of natural or synthetic compounds through rigorous forming and high-temperature sintering. Ceramics production is complex and each process may generate defects that affect the performance of ceramics. Presently, the inspection methods of ceramics are mainly divided into manual inspection and non-destructive inspection [19; 18; 12; 14].

This study focus on acoustic inspection of ceramics tableware. Acoustic non-destructive inspection, or acoustic emission, consists in applying stress to the piece and capturing the resulting sounds. However, the frailty of tableware does not allow the usage of such methods [3; 20]. Moreover, the manual inspection is still performed in many factories, where a subset of the production batch is inspected by trained personnel for defect detection. The procedure comprises a visual inspection and an acoustics inspection of ceramics. Therefore, it is prone to human error which is far from being the optimal solution. Additionally, ceramics are manufactured in mass production lines, thus manual inspection becomes a limiting factor to the speed of production. This drawback of manual inspection has costly consequences, e.g.

possible waste of materials, degraded quality of the shipped product, increased of labor time, which justify the need for an automated inspection and defect detection in ceramics.

The method of impulse noise uses a pendulum mechanic system to consistently produce a sound out of the ceramics material after the pendulum impact [4]. The audio signal is recorded and then analyzed. The authors of [4; 11] conducted time-frequency analyzes, i.e. determining audio signals' spectrogram [4] and wavelet transform [11], to compare cracked and uncracked plates' impact sounds. The results showed noticeable differences between cracked and uncracked plates, with cracked plates presenting less content concerning the lowest and highest frequency bands than uncracked plates. Latterly, the authors used an Artificial Neural Network to classify ceramics plates using audio signals' spectrum to train and test the model [5]. Recently, the authors used a Convolutional Neural Network to classify cracked and uncracked plates [16]. Images of impulse noise graphics were used to train and test the model and results showed a slightly higher accuracy than in [5]. The authors successfully classified cracked plates using deep learning models, but the dataset is too small to rely on or to generalize the obtained results.

Furthermore, other measurements were explored to analyze impulse audio signals in the time-frequency domain and compare cracked and uncracked plates data. The authors of [2] computed and analyzed audio signals Wigner-Ville distributions (i.e. quadratic/bilinear time-frequency distribution); conducted high-order spectral analysis (i.e. bispectrum and trispectrum) to obtain extra phase-related information; and calculated audio signals' mean value and peak root mean squared statistical measures. These features were very good in distinguishing between uncracked and cracked ceramic plates. The same authors compared bi-coherence within and/or between audio signals from uncracked and cracked ceramic plates. The results show that a durable plate displays a high magnitude peak in bi-coherence data, while a cracked plate has more than one peak with low magnitudes. Therefore, bi-coherence can be used in distinguishing the state of durability and crackness of the ceramic plates [1]. The authors of [15] calculated the coherence and transfer functions between uncracked and cracked plates, showing that there is less correlation between uncracked and cracked plates within lower frequencies.

The research on automated acoustic inspection of tableware ceramics is very limited. Generally, it focuses on either comparing time-frequency domain measurements calculated from audio signals of uncracked and cracked plates, or classifying ceramic plates based on Neural Networks. While, in the former, measurements inform a simple algorithm of differences between and/or within uncracked and cracked plates to distinguish them; in the latter, a deep learning model is trained and tested on very little data, thus, the results can not be generalized.

In this work, we propose an uncertainty aware automated acoustic inspection using a supervised machine learning model based on a set of novel acoustic features to classify ceramic plates of the same type, as cracked and uncracked. In this long paper, we present the results of experiments conducted on a dataset collected within the laboratory and using the impulse noise method. The main contributions of this paper are: 1) description of 192 features selected for the acoustic inspection of ceramic plates; 2) comparison of model calibration results regarding three different classifiers; 3) study of different sources of uncertainty for classification with rejection option through uncertainty quantification measures.

The rest of the manuscript is organized as follows: Section 2 describes the acquisition setup, protocol, and dataset preprocessing methods; Section 3 details the features extracted to the acoustic inspection; Section 4 depicts machine learning and uncertainty quantification methods; Section 5 presents and discusses the study results; and Section 6 indicates study conclusions and outlines future research course.

2. DATA COLLECTION AND PREPROCESSING

An impact pendulum was used to produce similarly reproducible impacts on a ceramic plate and sounds out of it, i.e. method of the impulse noise. The hammer attached to the end of the impact pendulum was made of steel covered with plastic tape, not damaging the ceramic plate during the impact. The impact actuator was constructed by the research team that ensured the hammer's initial angle θ position before each impact and the impact pendulum length l are constant over the data collection (see Fig. 1).

Fig. 1 displays the components of the data collection setup. Each trial consisted of: 1) starting recording within the computer audio software (i.e. Audacity); 2) waiting around 3 seconds while collecting some background noise; 3) applying an impact to the piece using the impact actuator; 4) recording the impact sound for approximately 1.5 seconds, enough to guarantee it fades out; 5) rotate the plate either 45°, or 135°, in either direction and repeat the steps 3), 4) and 5) three more times; 6) stop the recording within the computer's audio software. Data was collected at 48 kHz with a bit depth of 32 bits. We choose to collect more than one impact sound per plate because, if the plate has discontinuities in its medium (e.g. structural defects), the impact sound will depend on the place of the impact in the plate. Therefore, only one impact per plate could not be informative enough.

Regarding the setup components, we used the impact actuator to generate impacts on a ceramic plate to produce audible sounds. We settled the plate onto a rubber mat to maximize the duration of the sound. We used the microphone Sennheiser MKE300 to collect sound and the audio interface Focusrite Clarett 2Pre to convert microphone analog audio signals to digital. Lastly, the interface was connected to the computer to transmit digital audio signals and record them into the computer using the audio software.

We collected data in a total of 31 ceramic plates of the same type, 16 cracked and 15 uncracked. The audio dataset consists of a group of samples, each including the piece ID and label (i.e. cracked or uncracked plate), and an impact sound. Note that it is expected to have four samples for each combination of piece ID and label, corresponding to the expected number of impacts. Therefore, we had a total of 124 samples (i.e. 31×4).

The following detailed algorithmic methods were conducted using Python 3.7. To gather the audio dataset, several preprocessing methods were performed. Firstly, data was loaded from individual trials' audio files; within the data in each audio file, the onsets of the interesting events (i.e. impact sounds) were detected and the events segmented, generating four samples per audio file. At this stage, the

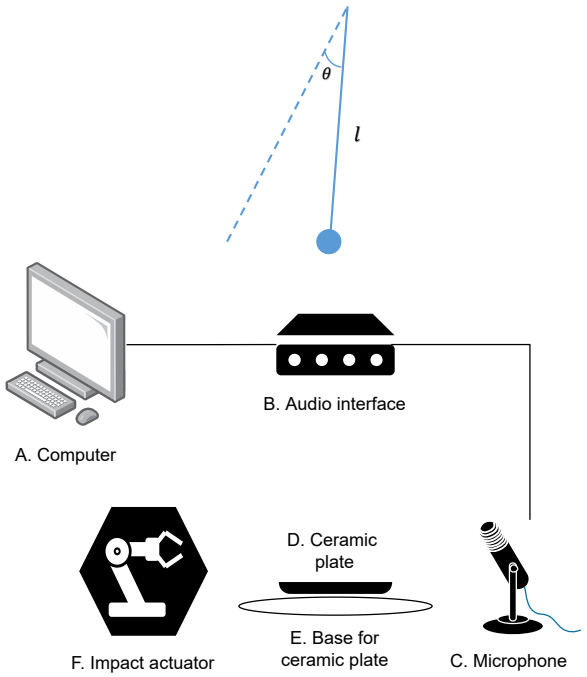


Figure 1: Impact actuator (top) - where θ defines the hammer's initial position before each impact and l is the pendulum length. Audio signals data collecting setup system (bottom) - it is composed by: A. computer (end-processing and storage equipment); B. audio interface (converts microphone's analog signals into a digital format the computer and audio software recognize); C. microphone (audio collecting equipment); D. ceramic plate; E. base for the ceramic plate (supports the ceramic plate during impacts); and F. impact actuator (generates impacts on a ceramic plate to produce audible sounds).

initially recorded background noise was also segmented (i.e. three first seconds of the recording).

In terms of data quality, the lack of it can result in noise, saturation (i.e. hardware-related problem due to electrical components "overloading"), or distortion (i.e. if audio signals are saturated and there is an alteration of the waveform state and shape). Hardware-related data quality issues were overlooked as solving them is out of the scope of this study. Nonetheless, these were not observed and were considered to be unlikely to happen within our laboratory settings. Secondly, for each sample, the background noise and the impact sound data were used to evaluate audio data quality in terms of noise. We calculated the sample's Signal-to-Noise ratio (SNR); if SNR value was below the threshold of 8 dB, we dropped the sample, otherwise, we kept it. After the data quality check procedure, 116 samples remained to further analysis, which means that data from two ceramic plates, both uncracked, was excluded (i.e. (124-116)/4).

Thirdly, we removed the laboratory background noise from the impact sound using spectral subtracting (i.e. subtracts the spectrum of the segmented background noise, which was captured at the beginning of the recording, to the spectrum of the segmented impact sound). Fourthly, we collected 30 minutes of background noise within a ceramic plate production line and use it to augment our dataset. The augmentation consisted of adding twenty-five times a randomly selected segment of "real" factory background noise to a sample's impact sound, which augments each sample to twenty-five samples. Afterwards, we obtained 2900 augmented samples (i.e. 116 × 25), 1600 cracked and 1300 uncracked plates (16 × 4 × 25 and (15-2) × 4 × 25, respectively).

3. CERAMICS INSPECTION ACOUSTIC FEATURES

The acoustic features were extracted for each impact audio data and for its initial portion (i.e. excludes the impact damping segment). The features were divided into two domains, spectral and temporal. Spectral features required the determination of audio signals' short-time Fourier transform, while temporal features were computed from the base audio time series. We computed a total of 192 features per sample - 2 signal's portions × [16 single-valued features + (4 multiple-valued features × 20 coefficients)]. Therefore, after feature extraction, we obtained a data frame of 2900 (samples) rows × 192 (features) columns.

Features choice for the acoustic inspection of ceramics was performed based on the literature review and the authors' visual inspection of impact audio data. These features will be described in the following sections.

3.1 Spectral

Frequency domain decay constants were calculated to capture the impact's exponential decrease over time.

3.1.1 Frequency decay

Exponential decay constant value in time of the magnitude to a given frequency band of interest (Eq. 1, where x is the time and y is the spectrum magnitude of a frequency band).

3.1.2 Spectral decay

Exponential decay constant value in frequency of the magnitude (Eq. 1, where x is the frequency and y is the spectrum magnitude).

$$-\frac{\ln y}{x} \qquad [\text{s}] \quad (1)$$

3.1.3 Spectral flatness

Ratio between the geometric mean and the arithmetic mean of the energy spectrum (Eq. 2, where x is the spectrum magnitude). Spectral flatness quantifies how much a sound resembles a pure tone, as opposed to being noise-like. Values can vary between 0 and 1. A high value indicates that the spectrum has a similar amount of power in all spectral bands (i.e. similar to white noise), while a low value indicates that the spectral power is concentrated in a relatively small number of bands. This feature is dimensionless (dn).

$$\frac{\sqrt[N]{\prod_{n=0}^{N-1} x_n}}{\frac{\sum_{n=0}^{N-1} x_n}{N}} = \frac{e^{\frac{1}{N}\sum_{n=0}^{N-1} \ln x_n}}{\frac{1}{N}\sum_{n=0}^{N-1} x_n} \qquad [\text{dn}] \quad (2)$$

3.1.4 Spectral energy

Sum of squared spectrum magnitude (Eq. 3, where x is the spectrum magnitude and N is the length of the signal).

$$\sum_{n=0}^{N-1} |x_n|^2 \quad [\mathrm{V}^2] \quad (3)$$

3.1.5 High frequency content

Sum of the spectral magnitudes of bins multiplied by each bin "position" (Eq. 4, where x is the spectrum magnitude and N is the length of the signal). It measures the amount of high-frequency content in a signal.

$$\sum_{n=0}^{N-1} n|x_n| \quad [\mathrm{V}] \quad (4)$$

3.1.6 Harmonic product spectrum derived features

The harmonic product spectrum is determined by downsampling the spectrum a given number of times and multiplying the downsampled spectrums. The harmonic product spectrum was used to extract the spectrum's fundamental frequency, relative harmonics ratio, the harmonic centroid and energy, the noisiness and odd-to-even ratio:

3.1.7 Fundamental frequency

If an audio signal is a musical note, its spectrum consists of a series of peaks corresponding to harmonic components at integer multiples of the signal's fundamental frequency. In this case, the harmonic product spectrum displays a clear maximum peak at the fundamental frequency. This is expressed in Hz.

3.1.8 Relative harmonics ratio

Ratio between the sum of harmonic product spectrum's relative harmonics power and the fundamental frequency power. The relative harmonics were defined as the harmonic product spectrum's peaks, excluding the peak at the fundamental frequency. This feature is dn.

3.1.9 Harmonic centroid

A variation on spectral centroid based upon frequency instead of bins (Eq. 5, where x is the magnitude of the harmonic product spectrum, f is the corresponding frequency and N is the length of the signal).

$$\frac{\sum_{n=0}^{N-1} x_n f_n}{\sum_{n=0}^{N-1} x_n} \quad [\mathrm{Hz}] \quad (5)$$

3.1.10 Harmonic energy

Sum of squared harmonic product spectrum magnitude (Eq. 3, where x is the harmonic product spectrum magnitude and N is the length of the signal).

3.1.11 Noisiness

The difference between the spectral energy and the harmonic energy. It represents how noisy a signal is. This is expressed in V^2.

3.1.12 Odd-to-even ratio

Odd-to-even ratio among the harmonics of an audio signal. Its value is higher for sounds with predominantly odd harmonics and lower for sounds with predominantly even harmonics (Eq. 6, where x is the harmonic product spectrum and N is the length of the signal).

$$\frac{\sum_{n=0}^{(N-1)/2} |x_{2n+1}|^2}{\sum_{n=0}^{(N-1)/2} |x_{2n}|^2} \quad [\mathrm{dn}] \quad (6)$$

3.1.13 Mel-frequency cepstrum and its related features

The signal's **cepstrum** is the inverse Fourier transform of the logarithm of its spectrum; it is used to explore periodic structures in frequency spectra. Mel-frequency cepstrum is a transformation of the power spectrum on a nonlinear Mel scale of frequency. **Mel-frequency cepstral coefficients** are coefficients that collectively form the Mel-frequency cepstrum. These coefficients provide information related to human auditory perception. In this work, we decided to determine 20 coefficients. **Mel-frequency cepstral coefficients delta and delta-delta** were also determined (Eqs. 7 and 8) as delta features capture sound differences in time.

$$\Delta_{MFCC_i} = MFCC_i - MFCC_{i+1} \quad [\mathrm{dn}] \quad (7)$$

$$\Delta\Delta_{MFCC_i} = \Delta_{MFCC_i} - \Delta MFCC_{i+1} \quad [\mathrm{dn}] \quad (8)$$

3.1.14 Linear prediction cepstrum and its related features

Linear predicted cepstrum uses a linear combination that predicts the current value of the cepstrum based on past samples. **Linear predicted cepstral coefficients** are coefficients that collectively form the linear predicted cepstrum. In this work, we decided to determine 20 coefficients. We also calculated the **cepstral peak prominence** which is the difference between the linear predicted and the "real" first peak of the signal's linearly predicted cepstrum's magnitude.

3.1.15 Period

The frequency corresponding to the higher peak in the signal's spectrum envelope.

3.2 Temporal

3.2.1 Harmonic to noise ratio

Ratio between the maximum value of the normalized autocorrelation of the signal and one minus that maximum value. This is expressed in dB.

3.2.2 Jitter (relative)

Relative variation in signal's frequency in percentage (Eq. 9, where T are the extracted period lengths and N the length is the number of extracted periods).

$$\frac{\frac{1}{N-1}\sum_{n=0}^{N-2} |T_n - T_{n+1}|}{\frac{1}{N}\sum_{n=0}^{N-1} |T_n|} \times 100 \quad [\%] \quad (9)$$

3.2.3 Shimmer (absolute)

Absolute variation in signal's amplitude in dB (Eq. 10, where A are the extracted peak-to-peak amplitudes and N the length is the number of extracted periods).

$$\frac{1}{N-1}\sum_{n=0}^{N-2} |20 \log A_{n+1}/A_n| \quad [\mathrm{dB}] \quad (10)$$

4. CERAMICS QUALITY CLASSIFICATION AND UNCERTAINTY QUANTIFICATION

A supervised machine learning model was constructed to automate the classification of cracked and uncracked ceramic plates. We divided our study into two experiments to check the impact of uncertainty quantification on the train and test sets classification performances using feature data (1) without and (2) with supervised selection.

4.1 Experiment 1

In the first place, we preprocessed the dataset: we encoded the samples' labels (cracked as 0 and uncracked as 1); scaled feature values from 0 to 1; and removed features with variance of 0 across samples for further analysis, only removing the first Mel-frequency cepstral coefficient.

We split the data into train and test sets in a proportion of 90% (train) and 10% (test). Notice that the number of samples between classes is unbalanced (i.e. 1600 cracked samples and 1300 uncracked samples), and that we had more than one sample per plate (i.e. 4×25). Thus, we split our data, stratifying it per plate's label and grouping it per plate's ID; this ensures that the proportion of cracked/uncracked samples is identical between both sets and that the samples derived from the same plate are not in the different sets (i.e. train and test sets), respectively.

Next, we fine-tuned three supervised machine learning algorithms: Gaussian Näive Bayes (Gaussian-NB), Random Forest (RF), and Support Vector Machines (SVM).

In real-world settings, the most accurate classification of ceramic plates is crucial to reliably reduce quality management efforts while accelerating automated and artificial intelligence-based instruments. Therefore, besides the performance evaluation, the use of uncertainty quantification methods for the classification with rejection option were explored in this work.

Before quantifying uncertainty, we performed several steps to improve and assess models' predicted probabilities and labels. Firstly, we calibrated models' predicted probabilities using the sigmoid function; we assessed the calibration by analyzing uncalibrated and calibrated models' performance metrics of calibration (i.e. brier loss and log loss) and classification (i.e. precision, recall, and f1-score) on the test set. Following, we performed a 5-fold cross-validation using the training set and predicted test set labels using the best model. In the former, we calculated accuracy within each validation fold and reported the average accuracy and its standard deviation to evaluate the classification performance; while, in the latter, we determined test accuracy, precision, recall, and f1-score.

Afterward, we applied different uncertainty quantification measures in order to evaluate their use for the classification with the rejection option setting. A distinction between two different sources of uncertainty is commonly made: aleatoric and epistemic uncertainty. Aleatoric uncertainty refers to the notion of randomness, and it is related to the data-measurement process. Epistemic uncertainty refers to the uncertainty associated with the model and the lack of knowledge [6]. The most well-known measure of uncertainty is the (Shannon) entropy of the predictive probabilities. This measure of uncertainty primarily captures the shape of the distribution and, hence, is mostly concerned with the aleatoric part of the overall uncertainty. Besides predictive entropy, we also explored the decomposition of aleatoric, epistemic, and total uncertainty by means of ensemble methods, as proposed in previous work [6; 8]. Since only RF is an ensemble method, for uncertainty quantification we designed a bootstrap approach using 10 bootstrap samples to create an ensemble for Gaussian-NB and SVM algorithms. Lastly, we evaluated the predictive uncertainty measures using accuracy-rejection curves and defined the optimal threshold for the rejection using the method proposed by Fisher et al [9], where the rejection cost was equal for both false negatives and a false positives (i.e. rejection cost was set to 0.5).

4.2 Experiment 2

In this experiment, we performed all steps described above in section 4.1 - Experiment 1 - and also used a supervised feature selection in the preprocessing stage. We used the feature selector based on percentile of the highest scores that calculates which features present significant difference in means between two independent populations (i.e. cracked and uncraked ceramic plates). In other words, it performs a hypothesis test for each feature; the lower is the p-value for each feature, the higher is the evidence that there is a difference between the populations means. We choose the 25^{th} percentile which corresponds to the percentage of features to keep.

5. RESULTS AND DISCUSSION

Table 1 presents the calibration results for both experiments. Observing experiment 1 results, model calibration showed a positive effect on predictions of Gaussian-NB and RF classifiers, while it had a little negative effect on SVM. In experiment 2 results, model calibration had a positive effect only on Gaussian-NB classifier predictions and had a negative effect on predictions of RF and SVM classifiers. Nevertheless, in general, the SVM classifier presented higher classification performance values for both experiments. Thus, we selected it, both uncalibrated and calibrated, for the subsequent steps.

Notice that the loss values are smaller and the classification performance measurements values are generally greater for experiment 2 than for experiment 1. These results were expected as in experiment 2 we applied the supervised feature selection.

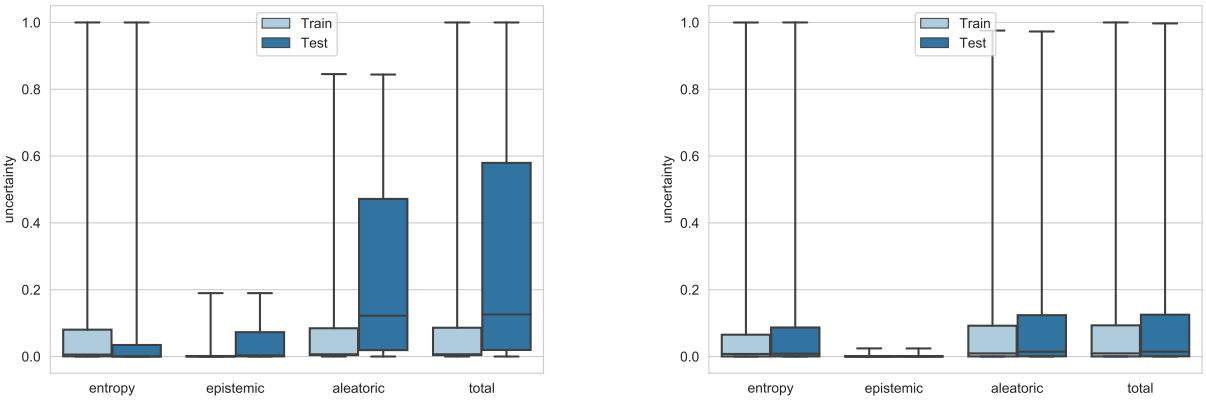
After performing a 5-fold cross-validation using SVM uncalibrated and calibrated classifiers separately, we obtained average accuracy values of $75 \pm 11\%$ and $96 \pm 5\%$ - for experiment 1 -, and $77 \pm 8\%$ and $98 \pm 2\%$ - for experiment 2 - respectively. Thus, we decided to conduct uncertainty quantification and uncertainty-based rejection of predictions using only the calibrated SVM classifier.

Furthermore, the prediction classification results on the test set, using the calibrated and validated SVM classifier, were: accuracy of 81%; f1-score of 78%; the precision of 85% and recall of 81% - for experiment 1 -; accuracy of 96%; f1-score of 96%; the precision of 96% and recall of 96% - for experiment 2.

The classification performance in train (i.e. cross-validation) was higher than in the test for both experiments, even though in experiment 2 the difference was smaller than in experiment 1. Hence, uncertainty quantification can help to detect the source of the decrease in test set classification performance. Fig. 2 shows the distributions of values for different

Table 1: Model calibration performance results for both experiments (Exp.1 and Exp.2). The calibration metrics used were: Brier loss and log loss; and the classification metrics used were: f1-score, precision and recall. *Fine-tuned model: RF (maximum depth of the tree of 8 and minimum number of samples required to split an internal node of 4) and SVM (regularization parameter of 100 and kernel coefficient of 0.01).

		Exp.1			Exp.2		
Metric	Model	Gaussian-NB	RF*	**SVM***	Gaussian-NB	RF*	**SVM***
Brier loss	Calibrated	0.14	0.11	**0.06**	0.082	0.147	**0.036**
	Uncalibrated	0.15	0.14	**0.05**	0.096	0.11	**0.037**
Log loss	Calibrated	0.52	0.32	**0.20**	0.276	0.481	**0.133**
	Uncalibrated	2.26	0.43	**0.18**	1.12	0.342	**0.133**
F1-score (%)	Calibrated	75	72	**88**	86	67	**93**
	Uncalibrated	75	77	**90**	86	81	**93**
Precision (%)	Calibrated	76	72	**99**	79	77	**98**
	Uncalibrated	75	65	**85**	79	70	**90**
Recall (%)	Calibrated	74	71	**80**	94	60	**89**
	Uncalibrated	75	94	**95**	96	99	**97**



(a) Exp.1 **(b)** Exp.2

Figure 2: Entropy, epistemic, aleatoric and total uncertainty values distributions with respect to train and test sets for both experiments.

types of uncertainty, concerning both the train and test, for both experiments. Since all measures are based on classical information-theoretic measures, each entropy value was normalized to the maximum entropy, i.e., the logarithm of the number of classes. Therefore, the uncertainty values are bounded between 0 and 1.

Comparing the train and test distributions for experiment 1 (in Figure 2.a), it is possible to see that the training set has low uncertainty values for all uncertainty measures (i.e. lower than 0.2). On the other hand, the test set shows higher values of median and interquartile range for aleatoric uncertainty, which results in higher values for total uncertainty. Comparing aleatoric uncertainty distributions with epistemic uncertainty distributions, we can reason that aleatoric factors are the strongest source of train and test classification performance differences. Although epistemic uncertainty distributions vary very little between the train and test, the test set has also more samples with higher uncertainty values. Thus, besides aleatoric processes, we can argue that our training set is not fully representative of our test set. Regarding predictive entropy, the uncertainty values distributions show opposite behavior.

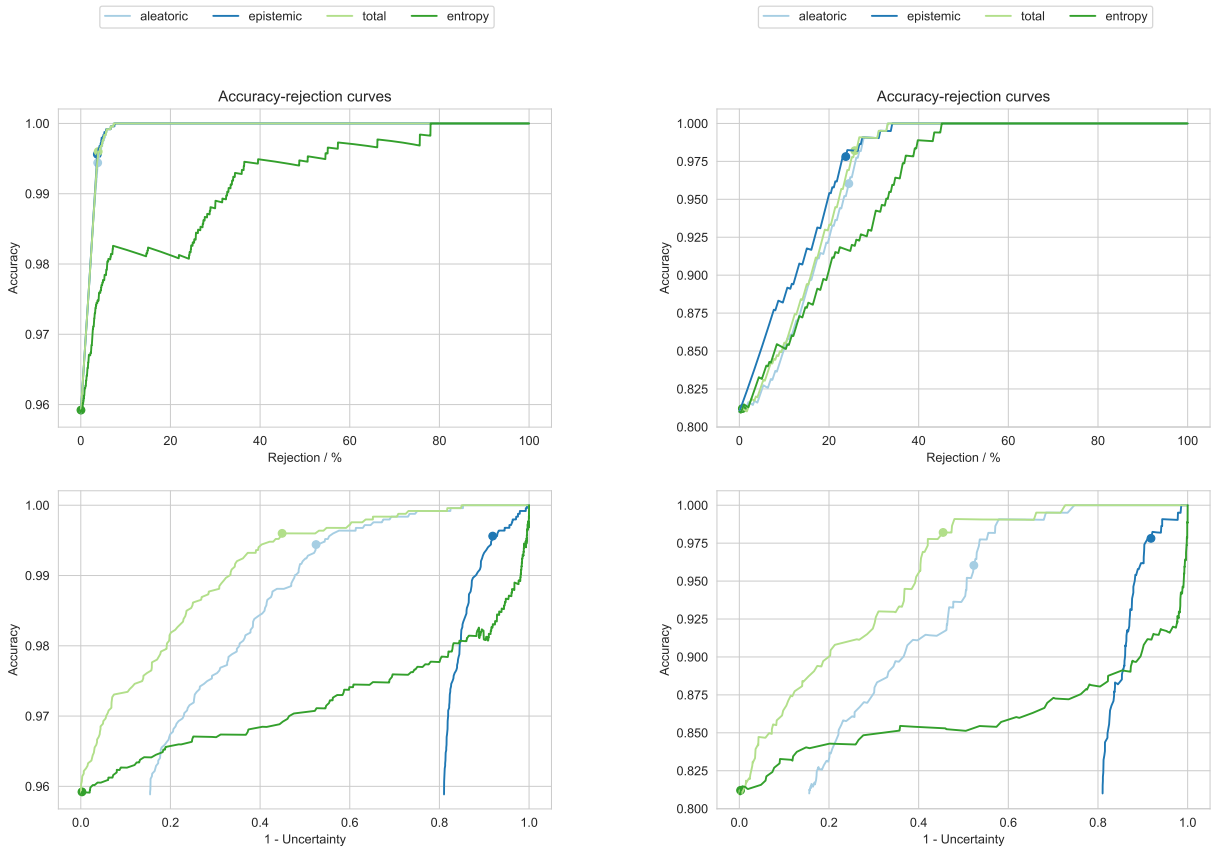
On the other hand, experiment 2 uncertainty quantification results (in Figure 2.b) show smaller differences between train and test sets distributions to the aleatoric and total uncertainty measures, while the median epistemic uncertainty equals to zero for both sets. These observations are consistent with the fact that the training and the test classification performance accuracy values are closer for experiment 2 than for experiment 1 results. Moreover, predictive entropy results differ from experiment 1, where test uncertainty values are greater than train uncertainty values, being in agreement with the other sources of uncertainty.

Detailed classification results with rejection for each uncertainty source are presented in Table 2.

The empirical evaluation of uncertainty quantification methods is not a trivial problem, due to the lack of ground truth uncertainty information. A common approach for indirectly evaluating the predicted uncertainty measures is using accuracy-rejection curves. Thus, Figs. 3 and 4 show experiments accuracy results for train and test, as a function of samples rejection percentage and one minus uncertainty.

Table 2: Optimal threshold values and classification results with rejection for each uncertainty source - aleatoric, epistemic, total and entropy.

Exp.	Uncertainty	Opt. Thres.	Acc. Train (%)	Rej. Train (%)	Acc. Test (%)	Rej. Test (%)
1	Aleatoric	0.525	99	3.8	97	25.0
	Epistemic	0.919	100	3.7	98	24.0
	Total	0.449	100	3.9	98	26.0
	Entropy	0.003	96	0.0	81	1.0
	W/out Rej.		96		81	
2	Aleatoric	0.127	99	1.1	98	5.3
	Epistemic	0.980	99	1.1	98	5.0
	Total	0.107	99	1.1	98	5.3
	Entropy	0.017	98	0.3	97	1.7
	W/out Rej.		98		96	



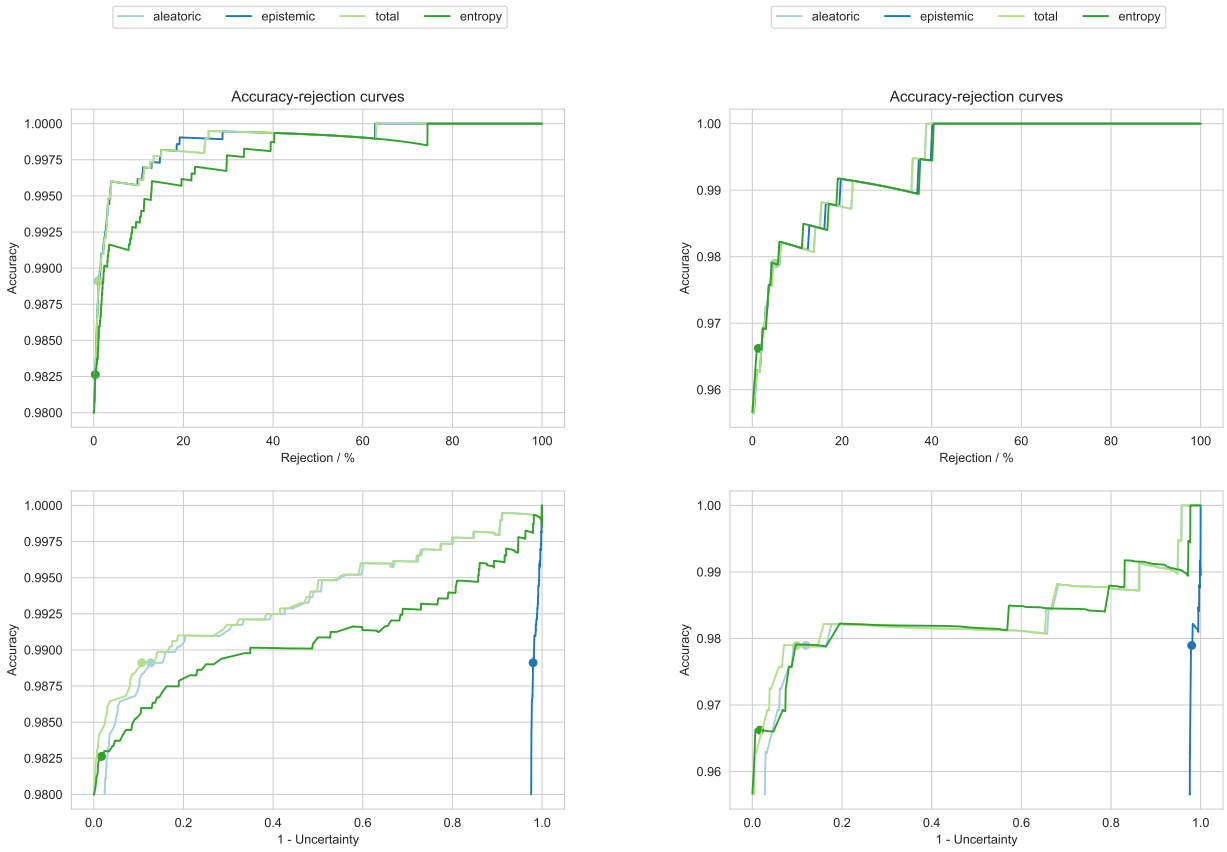
(a) Exp.1 - Train

(b) Exp.1 - Test

Figure 3: Accuracy as a function of samples rejection percentage (top) and one minus aleatoric, epistemic, total or predictive entropy uncertainty (bottom).

The rejection thresholds were defined using the global optimal threshold method from [9], where the difference between true and false rejects is maximized for the train uncertainty values interval. The rejection thresholds are represented Figs. 3 and 4 using a point colored with the same uncertainty source color; the exact values can be consulted in Table 2.

Analyzing Table 2, we note that the classification results with rejection using either aleatoric, epistemic, or total uncertainty are similar to each other for training and test sets and in consideration of both experiments results. Nonetheless, the percentage of rejected samples increases from train to test. We can visualize it by comparing Figs. 3.a) and 3.b), or Figs. 4.a) and 4.b), accuracy-rejection curves while spotting the rejection percentage for each uncertainty measure optimal threshold. On the other hand, the optimal threshold for predictive entropy is negligible for experiment 1, while for experiment 2 it has little effect on the accuracy



(a) Exp.2 - Train

(b) Exp.2 - Test

Figure 4: Accuracy as a function of samples rejection percentage (top) and one minus aleatoric, epistemic, total or predictive entropy uncertainty (bottom).

for train and test. Thus, the classification with rejection using entropy rejects none to very few samples for train and test, which reflects in a lack of better classification results than the baseline (i.e. without rejection in Table 2).

As the classification results with rejection using either aleatoric, epistemic, or total uncertainty are akin, we considered the classification with rejection based on total uncertainty.

Thus, in practice, from experiment 1 results, if the test set was a new batch of ceramic plates to be classified, the model could do it with 98% of accuracy, but it would not be able to classify 26% of the samples which would fall into a rejection class to further manual inspection. Additionally, selecting feature data resulted in an improved classification performance for train and mainly for test; then, from experiment 2 results, the same batch classification could reach the 98% of accuracy, but it would reject only 5.3% of the samples. Nonetheless, comparing experiment train results for classification with rejection option, there is a decrease of approximately 1% for experiment 2, but the percentage of rejected samples also decreased.

5.1 Limitations

The major limitations of our study are the usage of a rather small dataset collected within the laboratory, thus our model can fail on data acquired in "real" industrial settings, and the lack of labeling quality and reliability, as a few samples were classified as uncracked by the quality specialists, but those might have suffered structural damages in shipment from the factory to the laboratory.

6. CONCLUSION

Industry and Quality 4.0 pose the opportunity to integrate artificial intelligence-based technology into the quality management of products. In particular, quality control procedures of ceramics tableware require a demanding and faulty human manual inspection, both visual and acoustic. In this long paper, we focused on acoustic inspection, proposing an automated approach based on supervised machine learning methods to classify ceramic plates, as cracked and uncracked. We conducted experiments on a dataset collected within the laboratory, which was augmented using "real" factory background noise. Then, we extracted 192 implemented features per sample and split the resulting data into train/test sets in a proportion of 90/10. We selected three fine-tuned machine learning models (i.e. Gaussian-NB, RF, SVM) and assessed the classification and calibration performance of their train set predictions. Afterward, the SVM

calibrated classifier was selected as the best classifier, reaching 96% accuracy in the 5-fold cross-validation on the training set, and used in prediction on the test set reaching 81%. Additionally, in experiment 2, we also performed supervised feature selection. Furthermore, uncertainty quantification allowed us to reason that the major source of uncertainty is from the aleatoric part of the overall uncertainty. Moreover, when we conducted classification with rejection using the total uncertainty, train and test results significantly improved, reaching 100% and 98% accuracy for a rejection of 3.9% and 26% of samples - for experiment 1 - and 99% and 98% accuracy for a rejection of 1.1% and 5.3% of samples - for experiment 2 - respectively.

In future work, we intend to continue exploring classification with the rejection option through uncertainty quantification measures, namely improving the calibration procedure. Additionally, we will collect more data in different settings, improve the quality and reliability of the labeling, and add more information to the ceramic plate classification (e.g. defect type). Hereafter, the proposed automated approach can constitute a complementary method to accelerate, assist and improve quality control processes within ceramics tableware production lines.

7. REFERENCES

[1] O. Akgun. Damage detection in ceramic materials using bicoherence analysis. *Balkan Journal of Electrical and Computer Engineering*, page 300–306, 2020.

[2] O. Akgun. Spectral and statistical analysis for damage detection in ceramic materials. *Traitement du Signal*, 37(1):9–16, 2020.

[3] A. Y. Akimov. Methods of nondestructive control of ceramics (review). *Glass and Ceramics*, 47(6):213–217, 1990.

[4] T. Akinci. The defect detection in ceramic materials based on time-frequency analysis by using the method of impulse noise. *Archives of Acoustics*, 36(1):77–85, 2011.

[5] T. C. Akinci, H. S. Nogay, and O. Yilmaz. Application of artificial neural networks for defect detection in ceramic materials. *Archives of Acoustics*, 37(3):279–286, 2012.

[6] M. Barandas, D. Folgado, R. Santos, R. Simão, and H. Gamboa. Uncertainty-based rejection in machine learning: Implications for model development and interpretability. *Electronics*, 11(3):396, 2022.

[7] A. V. Carvalho, D. V. Enrique, A. Chouchene, and F. Charrua-Santos. Quality 4.0: An overview. *Procedia Computer Science*, 181:341–346, 2021.

[8] S. Depeweg, J. M. Hernández-Lobato, F. Doshi-Velez, and S. Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning, Jun 2018.

[9] L. Fischer, B. Hammer, and H. Wersing. Optimal local rejection for classifiers. *Neurocomputing*, 214:445–457, 2016.

[10] M. Javaid, A. Haleem, R. Pratap Singh, and R. Suman. Significance of quality 4.0 towards comprehensive enhancement in manufacturing sector. *Sensors International*, 2:100109, 2021.

[11] T. Kaynas, T. C. Akinci, O. Yilmaz, M. Ozgiray, and S. Seker. Defect detection for ceramic materials by continuous wavelet analysis, 2011.

[12] M. Kesharaju and R. Nagarajah. Particle swarm optimization approach to defect detection in armour ceramics. *Ultrasonics*, 75:124–131, 2017.

[13] D. P. Maganga and I. W. Taifa. Quality 4.0 conceptualisation: An emerging quality management concept for manufacturing industries. *The TQM Journal*, 2022.

[14] G. Morscher and Z. Han. Damage determination in ceramic composites subject to tensile fatigue using acoustic emission. *Materials*, 11(12):2477, 2018.

[15] A. Nayir. Coherence analysis and transfer function model for ceramic plate vibrations. *Journal of Vibroengineering*, 14:338–342, 03 2012.

[16] H. S. Nogay, T. C. Akinci, and M. Yilmaz. Detection of invisible cracks in ceramic materials using by pre-trained deep convolutional neural network. *Neural Computing and Applications*, 34(2):1423–1432, 2021.

[17] E. I. Papageorgiou, T. Theodosiou, G. Margetis, N. Dimitriou, P. Charalampous, D. Tzovaras, and I. Samakovlis. Short survey of artificial intelligent technologies for defect detection in manufacturing. *2021 12th International Conference on Information, Intelligence, Systems and Applications (IISA)*, 2021.

[18] T. Whitlow, E. Jones, and C. Przybyla. Failure prediction in ceramic composites using acoustic emission and digital image correlation. *AIP Conference Proceedings*, 2016.

[19] Z. Zhao. Review of non-destructive testing methods for defect detection of ceramics. *Ceramics International*, 47(4):4389–4397, 2021.

[20] Z. Zhao. Review of non-destructive testing methods for defect detection of ceramics. *Ceramics International*, 47(4):4389–4397, 2021.