Adaptive Shapley: Using Explainable AI with Large Datasets to Quantify the Impact of Arbitrary Error Sources

Birk Martin Magnussen*, Maik Jessulat†, Claudius Stern‡ and Bernhard Sick†
*biozoom services GmbH
34121 Kassel, Germany
e-mail: birk.magnussen@biozoom.net
†Intelligent Embedded Systems
Universität Kassel
34121 Kassel, Germany
email: {mjessulat, bsick}@uni-kassel.de
‡FOM Hochschule für Oekonomie & Management
Hochschulzentrum Kassel
34117 Kassel, Germany
e-mail: Claudius.Stern@fom.de

***Abstract*—Almost all sensors suffer from some level of uncertainty introduced from production inaccuracies. When the sensor data is processed by machine learning, quantifying the impact of such production inaccuracies on the output of the machine learning model becomes difficult.**

Certain neural network architectures, such as continuous feature networks, allow individual features and data to be omitted while still being able to correctly predict the result without the need for retraining. Such features can, for example, be individual channels of a sensor. This article proposes a method to use the capability to omit arbitrary features or sensor channels to calculate Shapley values for each sensor channel. These Shapley values represent the contribution of each individual channel to the measurement. They are defined using an arbitrary function called the "value function". If the value function is defined as the error of the current measurement, the Shapley values will represent the contribution of each sensor channel to the error of the measurement result.

By calculating Shapley values like this for a large unlabelled dataset of measurements, it is possible to understand how much measurement error was introduced by which channel of which sensor in each measurement. Averaging the Shapley values for each sensor in the dataset will then result in a metric for each channel of that sensor, which represents a contribution to measurement errors. By comparing these values to any arbitrary quality metrics for the sensor channels obtained in a calibration process or similar step, it is possible to correlate and quantify which value in the quality metric will cause how much of a measurement error, or whether the quality metric is even relevant for the measurement accuracy.

This article will show the efficacy and use case of the method on an example of the production and quality control of optical sensors based on multiple spatially resolved reflection spectroscopy.

***Index Terms*—explainable ai, big data mining, shapley values**

Supported by the state Hessen through funding as part of the Distr@l research grant 22_0041_2B.

I. Introduction

Optical sensors based on multiple spatially resolved reflection spectroscopy (MSRRS) consist of multiple light emitters and light detectors arranged in a matrix [1]. By using machine learning to compare the brightness obtained by the different detectors at different wavelengths when measuring a sample, it is possible to predict the concentration of carotenoids in the sample [2]. However, MSRRS-based sensors are highly dependent on an accurate calibration to counteract manufacturing inconsistencies. To understand and be able to quantify the impact of such manufacturing inconsistencies, quality criteria are collected during the calibration of a sensor. It is thus important to understand the relationship between the collected quality criteria and the actual resulting error in the carotenoid prediction for a sample.

Certain neural network architectures, such as continuous feature networks [3], have been developed that can predict the carotenoid concentration in a sample from MSRRS-based sensor data even when data from one or more light detectors is missing, without the need for retraining. This ability of the model to adapt to an input where the data from one or more detectors is omitted can be used to calculate so-called Shapley values [4] for a measurement. These Shapley values, calculated for each detector, represent the contribution of each detector to the "value" of the measurement prediction. By defining the value of a measurement prediction as the error of the current prediction, the Shapley values will represent the contribution of each detector to the error of the current measurement prediction.

In order to understand the impact of an arbitrary quality criterion available for each detector of each sensor, knowing the contribution to the prediction error for each detector of

each sensor is critical. By calculating the Shapley values for each detector for a large dataset of unlabelled measurements, it is possible to calculate an average contribution to the prediction error for each detector of each sensor in the dataset. The data-mined information on error contribution can then be correlated to the available quality criteria to understand whether any given quality criterion is relevant, and if so, what value of the quality criterion corresponds to what level of expected prediction error.

II. Related Work

Calculating the contribution of individual components to a whole can be approached with game theory, using so-called Shapley values. These represent the partial contribution of any player in a coalitional game to the outcome of the game [4].

Shapley values are a common tool in machine learning in order to explain model predictions. In this context, the Shapley values are commonly used to identify which feature supports the prediction in question, and which feature contradicts it. There are various approaches to this problem, with many implementations of the Shapley value for explaining model prediction slightly differing based on use-case or data structure [5].

A common issue Shapley value calculation faces is that since the evaluation effectively needs to be performed for the power set of features, calculating a Shapley value is NP-hard [6]. There are various strategies for either estimating the Shapley value based on assumptions about the context or calculating an approximation from only considering some instead of all possible subsets of features [7]. Of particular interest is an approach from the latter type of approximation, where feature permutations are stochastically sampled with a specified distribution [8], which makes it usable in many applications without the need for domain-specific assumptions.

A problem specific to the usage of Shapley values in machine learning is that in order to calculate Shapley values, players, or features in the machine learning context, need to be omitted from evaluation. However, most neural network architectures are unable to omit features directly. Different approaches use different definitions of removing values [9]. Common strategies include replacing a value to be omitted with a baseline value for the entire dataset [5] or replacing a value to be omitted with a value common for similar data points [10].

Several neural network architectures exist for processing input data of arbitrary size, including for sequential data [11], for point clouds [12], as well as for data with position-dependent features such as MSRRS-based sensor data [3]. However, so far, no research has used their properties to process input with some data omitted to calculate Shapley values. As choosing an unsuitable removal strategy has a negative effect on the accuracy of the Shapley value, making use of the adaptive properties of these neural network architectures for calculating the Shapley value is investigated in this article.

So far, most research on Shapley values focuses on explaining model predictions, rather than explaining prediction errors or related metrics. One exception is an analysis specific for calculating Shapley values for the uncertainty of a neural network for classification task, based on entropy [13]. Another article proposes using Shapley values from the loss of a neural network to classify error sources [14], however, as that needs many labeled data points, it is often difficult to put to use in practice. No specific research seems to have been done on using Shapley values to explain model prediction error in regression tasks for unlabelled data.

III. Shapley Values from Adaptive Neural Networks

Shapley values are originally defined by the following equation, where N is the set of all features, T is an arbitrary subset of N, i is the feature for which the Shapley value should be calculated, and $v(X), X \subseteq N$ is an arbitrary value function that assigns a value to the coalition of all features in X [4]:

$$\varphi_i(v) = \sum_{T \subseteq N \setminus \{i\}} \frac{|T|!\,(|N| - |T| - 1)!}{|N|!} \left(v(T \cup \{i\}) - v(T)\right) \tag{1}$$

In practice, this means that a Shapley value gives the average value addition of a feature for all possible permutations of all other features. However, as all permutations of one combination are expected to be identical, it is represented as an average of the value addition of a feature for all possible combinations weighted by the number of permutations per combination.

In the use case with continuous feature networks, a problem can be observed. The Shapley value, as defined in (1), requires the value of all possible combinations of features to be calculated. However, for predicting carotenoid concentration from MSRRS-based data, continuous feature networks only yield acceptable results if at least 72.5% of the data is still available [3]. Let m be the number of features that are required to calculate an acceptable result. It is possible to rewrite (1) so that instead of calculating the average over all possible permutations of features, we only calculate the average over all possible permutations of at least m features. For this, the weighting term of the sum needs to divide the weight not by the total number of permutations, given by $|N|!$, but by the number of permutations with at least m features, given by $(|N| - m)\,(|N| - 1)!$:

$$\varphi'_{i,m}(v) = \sum_{T \subseteq N \setminus \{i\}, |T| \geq m} \left(\frac{|T|!\,(|N| - |T| - 1)!}{(|N| - m)\,(|N| - 1)!} \cdot \left(v(T \cup \{i\}) - v(T)\right) \right) \tag{2}$$

It is to note, however, that (2) now no longer represents the total contribution of the respective feature to the total value, but only the additional contribution to the value when at least m other features are already present. For the purposes of error impact quantification, this is acceptable, as the trained model cannot predict acceptable results with less than 72.5% of the

data anyway, so error contribution in that regime is of little interest.

In case the number of combinations with at least m features is too large to reasonably calculate, it is possible to use existing, model-agnostic techniques to approximate the Shapley value in conjunction with the proposed method. Possible techniques are for example stochastic methods to sample the subsets of features [8], as long as consideration is given to include the condition to only sample combinations with at least m features.

In order for the Shapley values yielded by (2) to be the contribution to the error of a given detector, the value function $v(X)$ needs to be defined as the prediction error when only the detectors in X are available. Let $\psi(X)$ be an adaptive neural network capable of calculating a measurement prediction from an arbitrary number of detectors, such as a continuous feature network trained to predict the carotenoid concentration in a sample from MSRRS-based data. Similarly, let g be the ground truth for the measurement, then $v(X)$ can be defined as follows:

$$v(X) := |g - \psi(X)| \tag{3}$$

Unfortunately, as the dataset used is unlabelled, g is not available. As a result, g must be approximated from the available data. As $\psi(N)$, the prediction of the continuous feature network using all data available, would be affected by any error from any detector, it cannot be used as an approximation for g.

It can be assumed that the majority of the detectors used in the sensors comprising the dataset are reasonably accurate and that detectors with a significant error contribution are rare. This assumption will be supported by findings in section IV. Under this assumption, if one detector of a sensor has a significant error contribution, the other detectors are likely to yield a reasonably accurate result on average. This can be used to approximate g as the prediction of the continuous feature network using all detectors except i when used to calculate the error contribution of detector i:

$$g_i \approx \psi(N \setminus \{i\}) \tag{4}$$

From this, a value function for calculating Shapley values can be as follows:

$$v_i(X) \approx |\psi(N \setminus \{i\}) - \psi(X)| \tag{5}$$

The Shapley value is a weighted sum over the difference of two points of the value function:

$$v_i(T \cup \{i\}) - v_i(T), T \subseteq N \setminus \{i\} \tag{6}$$

In the case that detector i introduces a prediction error, (6) yields the marginal contribution of detector i to the coalition of detectors T, since in that case g_i can be assumed to be accurate. In the case that a different detector introduces a prediction error, assuming that the prediction error of detectors is independent of one another, the error in the assumed ground truth is present approximately equally in both components of the difference. Thus, even though g_i is adversely affected, (6) is still a good approximation of the marginal contribution of detector i to the coalition of detectors T.

In the remainder of this article, (5) is used as the value function for the calculation of all Shapley values.

IV. Applying Shapley Values to Large Datasets

Calculating the Shapley values as defined in section III yields one set of Shapley values for each measurement. In order to gain knowledge of per-sensor production inaccuracies, outliers from individual measurements need to be accounted for. To achieve this, for each sensor, Shapley values are first calculated for all available measurements and then averaged for each detector. The resulting average contribution to the error for each detector of every sensor can be used for further analysis.

The available dataset contains approximately $2\,500$ sensors with a total of $5\,500\,000$ measurements. The measurements were filtered for successful measurements only, with roughly $4\,000\,000$ measurements contributing to the Shapley values.

While only using detectors that reduce the overall error is ideal, a certain level of error contribution may be acceptable. To verify what Shapley value corresponds to the highest acceptable level of error contribution, the error magnitude ρ_i of any detector for an arbitrary measurement is calculated as follows:

$$\rho_i := \frac{1}{|N| - 1} \left(\sum_{t \in N \setminus \{i\}} \psi(N \setminus \{t\}) \right) - \psi(N \setminus \{i\}) \tag{7}$$

The result of (7) represents the difference between the average prediction of $|N| - 1$ detectors containing the detector to be analyzed and the prediction of all detectors except the detector to be analyzed. By averaging this difference over multiple measurements, a very rough approximation of the error introduced by the detector is gained.

Figure 1 shows a distribution of the error magnitude over the Shapley value for every detector in the dataset of $2\,500$

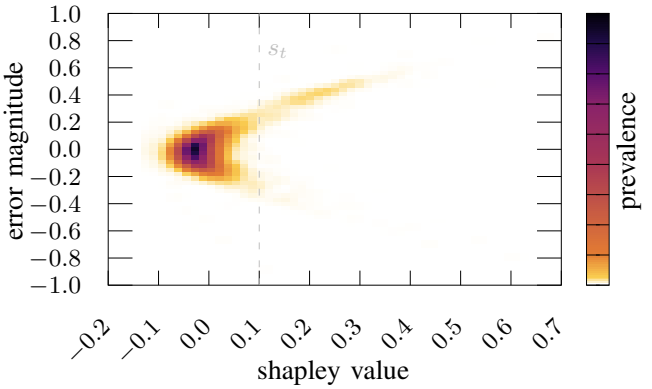

Fig. 1. The distribution of the error magnitude over the Shapley value of all detectors. s_t shows the threshold Shapley value.

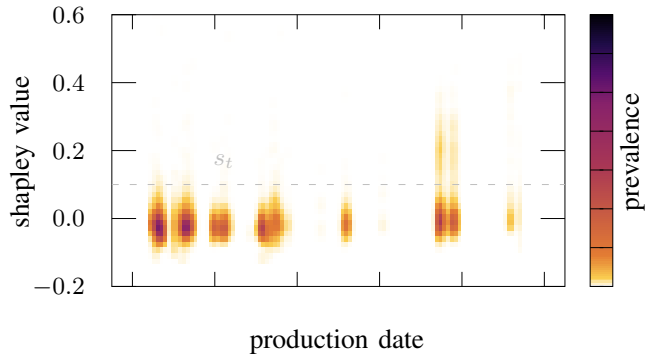

Fig. 2. The distribution of the Shapley value of all detectors over the production date, indicating differences between production lots. A systematic deviation in the second to last production lot can be seen.

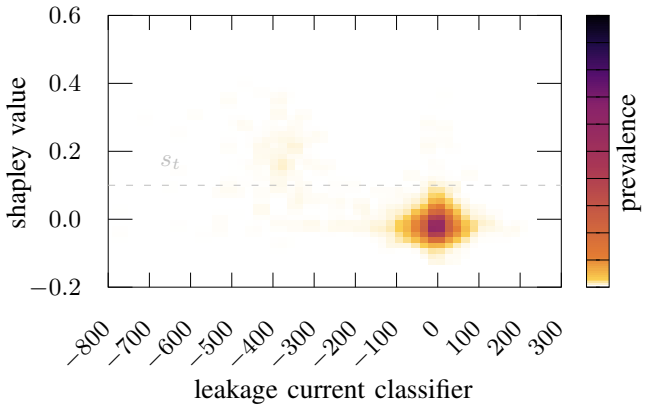

Fig. 3. The distribution of the Shapley value over the observed leakage current classifier (negative values indicate a high leakage current) of all detectors.

sensors. As the target for the maximum prediction error for this type of MSRRS-based sensor is ± 0.25, the maximum accepted Shapley value s_t is defined as 0.1 from this data. As a result, all detectors with a Shapley value exceeding the threshold of $s_t = 0.1$ are deemed of insufficient quality. In addition, Figure 1 shows that the overwhelming majority of detectors are accurate and do not introduce significant error. 96.7% of detectors have a Shapley value below the threshold of 0.1. Similarly, 71.6% of detectors have a Shapley value below 0, which means that the detector will generally improve the measurement accuracy when at least m other detectors are available. This supports the assumption about rare detectors with a significant error contribution from section III.

In order to validate the usefulness of the Shapley value as a quality metric, the distribution of Shapley values over production lots will be analyzed. Due to known and otherwise compensated production issues, one detector showed systematic deviations in the second to last production lot. Figure 2 shows that in the affected production lot, a significant increase of detectors with a higher Shapley value (and one above the threshold s_t) is observable. This shows that certain systematic deviations of the affected detectors are detectable with the presented Shapley value-based analysis method.

V. Shapley Value Interpretation

In order to use the calculated Shapley values to quantify the efficacy of arbitrary quality criteria, the distribution of Shapley values over any given quality criterion needs to be analyzed. For MSRRS-based sensors, one quality criterion of interest is the leakage current of the light detectors. This metric is expected to be relevant, as leakage current will cause a detector to not be able to differentiate between very low light and total darkness, resulting in loss of information. As the leakage current classifier can be measured easily during calibration, being able to use it as a calibration quality criterion is desired.

Figure 3 shows a heatmap of the distribution of Shapley values over a metric that measures the leakage current observed from the light detectors in MSRRS-based sensors,

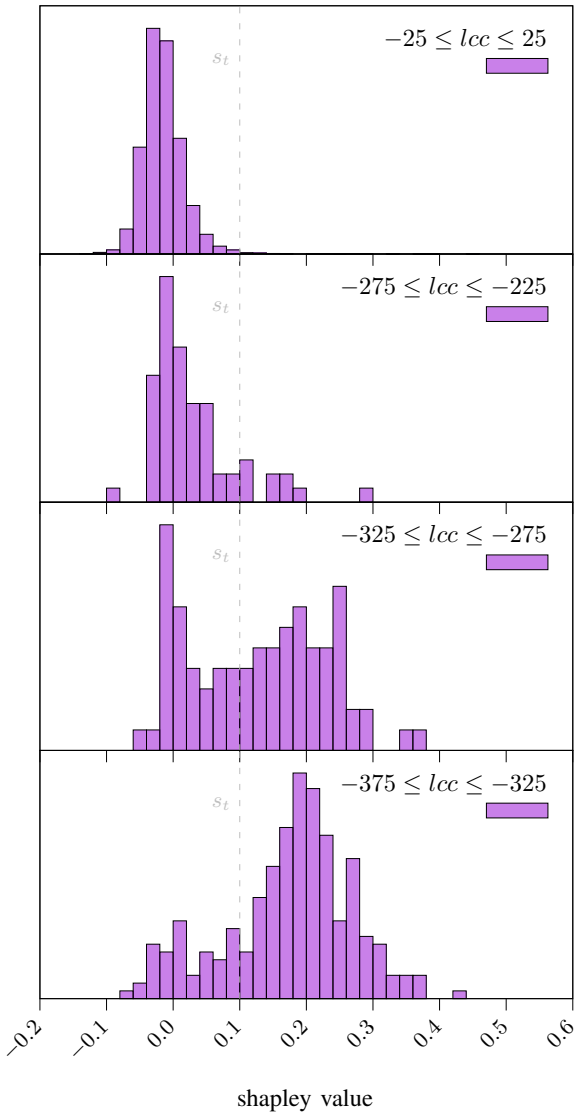

Fig. 4. Histograms of the relative distribution of Shapley values for detectors with different leakage current classifiers.

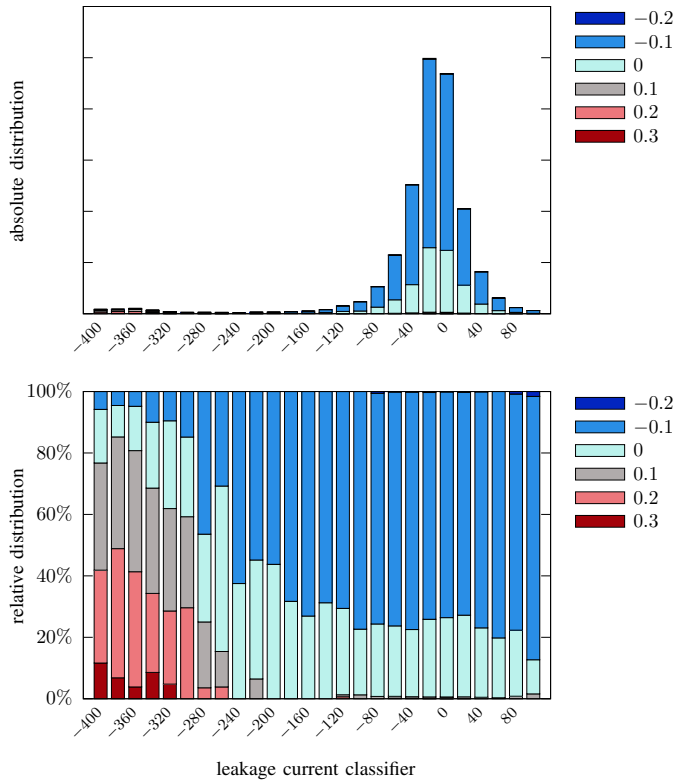

Fig. 5. The rough distribution of Shapley values over different leakage current classifiers. The top graph shows the absolute distribution, while the bottom graph shows the relative distribution. A shift in distribution is observable at a leakage current classifier of around -260.

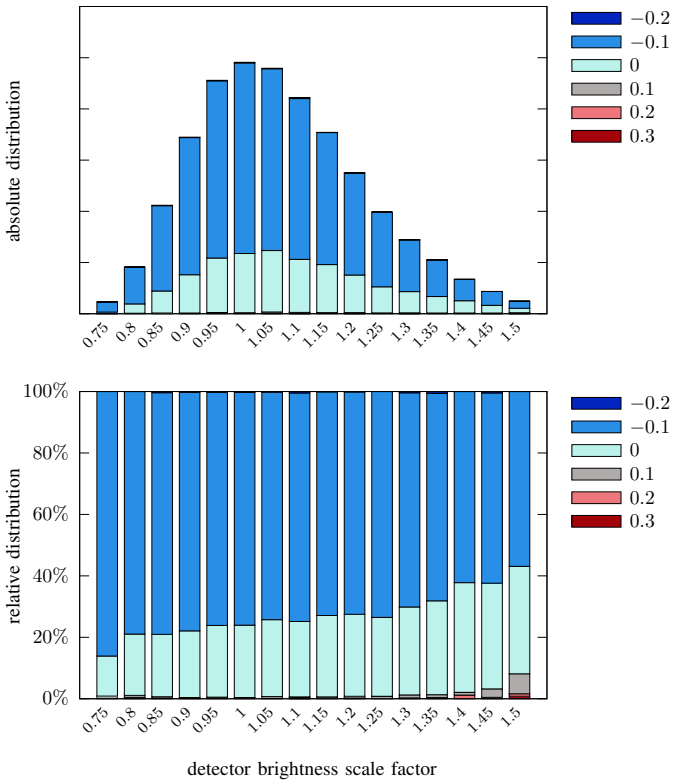

Fig. 6. The rough distribution of Shapley values over different detector brightness scale factors. The top graph shows the absolute distribution, while the bottom graph shows the relative distribution. No clear shift in distribution is observable.

where negative values indicate a higher leakage current. The first observation from Figure 3 is that the vast majority of detectors have a leakage current classifier close to 0, and thus no significant deviance in observed leakage current to the expected value. The second observation is that the detectors that do not have a leakage current close to 0 also have a much wider distribution of Shapley values.

Figure 4 helps to get a more detailed picture of how the Shapley values are distributed for different leakage current classifiers. The first histogram shows the distribution for leakage current classifiers close to 0, which shows that almost all detectors in this regime have a Shapley value below the threshold of 0.1. When looking at detectors with a lower leakage current classifier, it becomes apparent that the distribution changes to a majority of detectors exceeding the threshold Shapley value of 0.1.

In order to learn at which leakage current classifier the Shapley value distribution exceeds an acceptable level of detectors above the Shapley value threshold, an analysis of the Shapley value distribution over the leakage current classifier, as seen in Figure 5, can be used. As visible in Figure 5, there is a threshold at a leakage current classifier of around -260 where the rate of detectors with a Shapley value of 0.1 and higher increases sharply. From this, it can be deduced that -260 is a good threshold at which to reject detectors during calibration. Setting this criterion to -260 would keep 96.2% of detectors. For approximately 1.2% of all detectors with an acceptable Shapley value, this criterion is a false positive. At the same time, this criterion alone is able to catch 80.1% of all detectors with an unacceptable Shapley value. Of the detectors that are filtered by this criterion, 68.6% have an unacceptable Shapley value.

Such an analysis, correlating calibration quality criteria with Shapley values, can be performed for various other quality criteria as well. One notable example can be seen in Figure 6. The figure shows how much a detector's overall brightness is scaled from the calibration process. A high value means that the detector returns comparably dark results and needs to be scaled up, while a low value means that a detector returns bright results and needs to be scaled down. Unlike the distribution seen in Figure 5, the distribution in Figure 6 does not show any clear thresholds or any regions of the scale factor in which detectors with an unacceptably high Shapley value are prevalent. This suggests that using the detector's brightness scale factor is not a suitable metric based on which to reject an MSRRS-based sensor. Nonetheless, darker detectors appear to perform slightly below bright detectors as measurement precision is lost. The trend in Figure 6 suggests that if even darker detectors exist, a maximum acceptable brightness scale factor might be observed through the proposed method.

VI. Conclusion

This article introduces a method on how to calculate Shapley values to explain model predictions by using adaptive neural network architectures capable of directly omitting features. Furthermore, a method on how to use these Shapley values to predict the contribution of any feature to the error of one measurement prediction is introduced.

By applying the introduced method to all measurements of a large unlabelled dataset, a way to gain knowledge about systematic error sources from features in the dataset is presented. Comparing the Shapley values yielded from this process to error estimations, it is shown how the gained knowledge can be used to understand error sources. Similarly, the Shapley values can be correlated with arbitrary quality criteria and other metrics of the features of the models to quantify the efficacy of any given quality criterion. Furthermore, the proposed method allows any thresholds in the metrics beyond which prediction error is introduced to become apparent.

The proposed method for error quantification is examined on the example of optical multiple spatially resolved reflection spectroscopy-based sensors. The method is validated on known data and then used to quantify two quality metrics as an example. For one of the metrics, the method was able to determine that within the range present in the produced sensors, no region of the metric corresponds to a significant prediction error, showing that the chosen quality metric is of low efficacy. For the other metric, the method was able to show high efficacy and identified a clear threshold beyond which the prediction accuracy of the measurement exceeds acceptable limits.

This shows that in conclusion, the proposed method is capable of analyzing large unlabelled datasets, identifying the efficacy of different metrics and quality criteria, and is able to detect concrete thresholds for usable metrics.

References

[1] B. Magnussen, C. Stern, and W. Köcher, "Vorrichtung und verfahren zur bestimmung einer konzentration in einer probe," European Patent EP 3 013 217 B1, 2016.

[2] M. E. Darvin, B. Magnussen, J. Lademann, and W. Köcher, "Multiple spatially resolved reflection spectroscopy for in vivo determination of carotenoids in human skin and blood," *Laser Physics Letters*, vol. 13, no. 9, p. 095601, Aug. 2016.

[3] B. M. Magnussen, C. Stern, and B. Sick, "Utilizing continuous kernels for processing irregularly and inconsistently sampled data with position-dependent features," in *Proceedings of The Nineteenth International Conference on Autonomic and Autonomous Systems*, C. Behn, Ed., IARIA. ThinkMind, Mar. 2023, pp. 49–53.

[4] L. S. Shapley, "Notes on the n-person game – II: The value of an n-person game," RAND Corporation, Tech. Rep. ATI 210720, Aug. 1951.

[5] M. Sundararajan and A. Najmi, "The many shapley values for model explanation," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, Jul. 2020, pp. 9269–9278.

[6] X. Deng and C. H. Papadimitriou, "On the complexity of cooperative solution concepts," *Mathematics of Operations Research*, vol. 19, no. 2, pp. 257–266, 1994.

[7] H. Chen, I. C. Covert, S. M. Lundberg, and S.-I. Lee, "Algorithms to estimate shapley value feature attributions," *Nature Machine Intelligence*, vol. 5, no. 6, pp. 590–601, Jun. 2023.

[8] J. Castro, D. Gómez, and J. Tejada, "Polynomial calculation of the shapley value based on sampling," *Computers & Operations Research*, vol. 36, no. 5, pp. 1726–1730, 2009, selected papers presented at the Tenth International Symposium on Locational Decisions (ISOLDE X).

[9] I. Covert, S. Lundberg, and S.-I. Lee, "Explaining by removing: A unified framework for model explanation," *Journal of Machine Learning Research*, vol. 22, no. 209, pp. 1–90, 2021.

[10] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 4768–4777.

[11] D. Romero, A. Kuzina, E. Bekkers, J. Tomczak, and M. Hoogendoorn, "Ckconv: Continuous kernel convolution for sequential data," in *ICLR*, 2022.

[12] W. Wu, Z. Qi, and L. Fuxin, "Pointconv: Deep convolutional networks on 3d point clouds," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9613–9622.

[13] D. S. Watson, J. O'Hara, N. Tax, R. Mudd, and I. Guy, "Explaining predictive uncertainty with information theoretic shapley values," 2023.

[14] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, Jan. 2020.