2025 IMEKO TC-6 International Conference on
Metrology and Digital Transformation - **M4DConf 2025**
Benevento, Italy, September 3-5, 2025

Approximative Bayesian approach for uncertainty evaluation in machine learning-based hardness measurement

Junnosuke Takai[1], Yukimi Tanaka[1], Masahiro Yoshioka[1], Katsuhiro Shirono[1]

[1] *Research Institute for Engineering Measurement, National Metrology Institute of Japan, National Institute of Advanced Industrial Science and Technology (AIST), AIST Tsukuba Central 3, Umezono 1-1-1, Tsukuba, Ibaraki 305-8563, Japan, Email:takai.jun@aist.go.jp*

***Abstract* – In recent years, the application of machine learning to the field of metrology has been increasingly explored. In the evaluation of measurement uncertainty using machine learning, it is generally considered necessary to evaluate it through combining two types of uncertainties: aleatoric uncertainty, which arises from randomness, and epistemic uncertainty, which arises from systematic factors. However, the methods for evaluating these uncertainties have not yet been established. In this study, we evaluate measurement uncertainty in the Vickers hardness measurement using a Convolutional Neural Network. For this evaluation, we employ Monte Carlo Batch Normalization as an approximation of a Bayesian Neural Network to evaluate epistemic uncertainty. As a result, it was found that a reasonable evaluation is possible for materials similar to those in the training data.**

I. INTRODUCTION

Machine learning (ML) is a technology that enables the construction of models capable of representing functions based on data, even when the analytical form of the function is unknown or highly complex. In recent years, it has been applied across a wide range of fields. In particular, deep learning, a class of machine learning methods that allows for the construction of highly flexible and complex models, has seen widespread adoption. In the field of metrology as well, a lot of studies employing deep learning techniques have been reported.

For example, one of the authors, Tanaka, together with co-authors, reported a study on the Vickers hardness measurement that employed a Convolutional Neural Network (CNN), a type of deep learning model [1]. The Vickers hardness is measured using the diagonal length of an indentation created when a pyramidal indenter is pressed into a material. When this diagonal length is measured manually from an image of the indentation, there are some challenges including time consuming tasks and human depending differences in the results. Furthermore, in cases where the diagonal length is determined automatically from image features such as gradients, accurate measurement may be difficult due to factors such as rough surfaces or cracks. Tanaka et al. demonstrated that the use of a CNN for diagonal length measurement enables high efficiency and high accuracy evaluation across a wide variety of materials.

However, the evaluation of uncertainty when using machine learning has not yet been fully established. Thompson et al. (2021) stated that "Traditional uncertainty evaluation of the kind systematized by the GUM framework makes the assumption that the measurement model is known. Uncertainty evaluation in ML deviates from this paradigm in one key respect". One of the essential requirements for ensuring metrological traceability of measurements is the evaluation of measurement uncertainty. Establishing metrological traceability is significant in conformity assessment and quality control. In the context of the hardness measurements, the evaluation of measurement uncertainty is also required to ensure traceability.

Regarding measurement uncertainty in the use of machine learning, Hüllermeier et al. (2021) proposed that the uncertainty in machine learning outputs can be decomposed into two types: aleatoric uncertainty and epistemic uncertainty. They suggest that by evaluating these two components, it becomes possible to evaluate measurement uncertainty even when machine learning is employed [3]. Aleatoric uncertainty reflects the randomness in the data, whereas epistemic uncertainty represents the uncertainty arising from the imperfection of the model. Several methods have been proposed to estimate these two types of uncertainties [2–4].

In this study, we report on the evaluation of uncertainty in the measurement of the Vickers hardness using a CNN. For the estimation of epistemic uncertainty, we apply Monte Carlo Batch Normalization (MCBN), one of the methods proposed in previous research [4]. The structure of this paper is as follows. In Section 2, we provide an overview of the two types of uncertainties in machine learning and describe the MCBN. Section 3 presents our experiments on hardness measurement and uncertainty

evaluation using the MCBN, along with the corresponding results. Section 4 discusses and interprets the results presented in Section 3. Finally, Section 5 provides a summary of this study.

II. UNCERTAINTY EVALUATION THROUGH MONTE CARLO BATCH NORMALIZATION

A. Aleatoric uncertainty and epistemic uncertainty

In this study, we consider the problem of supervised learning, where a measurement model is represented (trained) using known measurement data (training data), and predictions are made for an unknown estimate via the regression. It is assumed that the training data are independently and identically distributed (i.i.d.), and generated from an unknown probability density function P. Associating the measurement data with the probability density function P implies that, even if complete knowledge of P were available, some variation in the output estimate would remain. This irreducible uncertainty is referred to as aleatoric uncertainty. On the other hand, in practice, it is not possible to determine the probability density function P perfectly. The uncertainty arising from this incomplete knowledge is referred to as the epistemic uncertainty. According to Hüllermeier et al. [3], the epistemic uncertainty can be further decomposed into two components: model uncertainty, which reflects whether the machine learning model is capable of accurately representing the measurement model, and approximation uncertainty, which results from the model's limited approximation capability due to the finiteness of the data.

However, in the case of flexible models such as ones developed by deep learning, it is generally assumed that the measurement model could be represented with high accuracy, provided that enough data were available. Therefore, in this study, where a deep learning model is employed, only approximation uncertainty is treated as epistemic uncertainty.

For uncertainty evaluation in measurement using machine learning, it is necessary to evaluate both aleatoric and epistemic uncertainties.

B. Monte Carlo Batch Normalization

Prior to discussing the evaluation of uncertainty, we provide an overview of Batch Normalization (BN), a technique widely used in deep learning to improve training stability and accelerate convergence. In deep neural networks, as illustrated in Fig. 1, input values are propagated through multiple layers, where each layer applies nonlinear transformations at each node, eventually producing a predicted output. While deeper networks tend to have greater expressive capacity, they are also more prone to instability and inefficiency during training.

One possible approach to addressing these challenges is BN. In typical deep learning frameworks, training data are divided into small subsets called mini-batches, and model parameters are updated based on each mini-batch. BN normalizes the output of each node in a layer to have zero mean and unity variance based on the data of each mini-batch. During the prediction, the normalization is performed using the parameters optimized during training (see Fig. 1).

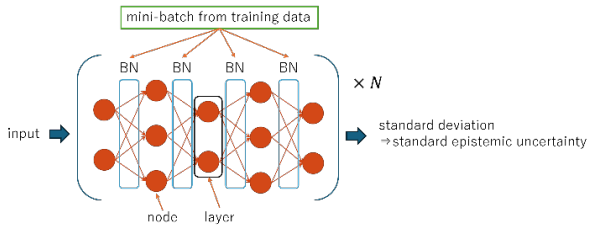


Fig. 1 Schematic diagram of the MCBN. N is the number of Monte Carlo sampling (sampling number of the mini-batches).

We then introduce Monte Carlo Batch Normalization (MCBN), proposed by Teye et al. [4], as a method for evaluating epistemic uncertainty. Several fundamentally different approaches have been proposed for evaluating epistemic uncertainty in the deep learning [2]. One such approach is the Bayesian Neural Network (BNN), which evaluates uncertainty by performing the Bayesian inference on the parameters of the underlying probability distribution P. In BNN, a predictive posterior distribution of the output estimates represents the epistemic uncertainty. However, due to the extremely large number of parameters in deep learning models, obtaining the predictive posterior distribution via direct Bayesian inference, such as that using Markov Chain Monte Carlo (MCMC) methods, is often computationally impractical.

To address this issue, Teye et al. proposed the MCBN as an approximate method for deriving the predictive posterior distribution. As mentioned previously, standard BN uses fixed normalization parameters during the prediction. In contrast, MCBN introduces stochasticity by randomly sampling mini-batches from the training data even during the prediction phase and applying the BN based on the sampled mini-batch. By repeating this sampling, multiple prediction estimates can be obtained for the same input data. According to Teye et al., the resulting distribution of the prediction estimates can be interpreted as an approximation of the predictive posterior in the BNN. In this study, we evaluate epistemic uncertainty by analyzing the distribution of prediction estimates generated through MCBN.

III. EXPERIMENTAL AND ANALYTICAL METHODS

This section provides an overview of the uncertainty analysis in the Vickers hardness measurement using CNNs. For details regarding the Vickers hardness measurement procedure, see [1].

The Vickers hardness is determined by measuring the lengths of the diagonals of the indentation formed when an indenter is pressed into a material. In [1], two types of CNNs are used to measure the diagonal lengths from images of the indentations. Each indentation image is captured at one of the following magnification levels:

$m = 10,\ 20,\ 40,\ 50,\ 100$. First, a coarse estimation of the four corner positions of the indentation is obtained using the first CNN. Then, around each estimated corner, a magnified image is extracted, so that each magnified image contains only one visible corner.

For each magnified image, a second CNN is used to predict the corner position more precisely, from which the diagonal length is computed. As the CNN outputs are in pixel units, the results are converted into physical lengths based on the corresponding magnification factor m.

Since the output of the first CNN does not significantly influence the final measurement, only the uncertainty associated with the second CNN is evaluated. Following [1], this CNN is referred to as the CNN-CP (CNN for the Corner Positions). The CNN-CP takes a 100×100-pixel image as input and consists of a 10-layer neural network. The activation function used is ReLU, and the BN is applied after each layer. The network is implemented using the PyTorch library, trained and optimized using the Adam optimizer with a smooth L1 loss function [5]. A mini-batch of 64 is used. Training is performed for the first 100 epochs with a learning rate of 10^{-4}, followed by an additional 200 epochs with a learning rate of 10^{-6}.

The training dataset consists of indentation images obtained from reference blocks with nominal Vickers hardness values of 200 HV, 300 HV, 400 HV, 500 HV, 600 HV, 700 HV, 800 HV, and 900 HV, as well as manually labeled indentation images of titanium dioxide (TiO_2). Further details are available in [1].

The standard epistemic uncertainty, u_{eps}, is estimated using the MCBN. As the measured value is defined as the average of the two diagonals, and each diagonal requires two corner predictions, a total of four CNN-CP outputs is needed for a prediction estimate. To account for possible correlations among the four outputs, the average diagonal length is calculated in each Monte Carlo sample. The standard deviation of these averages is then used to determine u_{eps} as follows:

$$u_{\mathrm{eps}}(x) = \sqrt{\frac{\sum_i \left(f_{\mathrm{MC},i}(x) - \bar{f}_{\mathrm{MC}}(x)\right)^2}{N-1}}, \qquad (1)$$

where x is the input image, $f_{\mathrm{MC},i}(x)$ is the average diagonal length predicted in the i-th Monte Carlo sample, and $\bar{f}_{\mathrm{MC}}(x)$ is the mean of all $N = 5000$ samples. Note that the standard epistemic uncertainty depends on the input.

The standard aleatoric uncertainty, u_{ale}, is derived from the training dataset. Since outliers may be present in the training data, a robust estimation method is employed. Specifically, the median absolute deviation (MAD) between the predicted estimates and the measurement value is calculated. Based on the properties of the normal distribution, the MAD is scaled by 1.4826 to obtain the standard uncertainty:

$$u_{\mathrm{ale}}(x) = c_m \times 1.4826 \times \mathrm{Med}\left(\left\{\left|g\left(x_{\mathrm{train},i}\right) - t_{\mathrm{train},i}\right|\right\}_{i \in T}\right) \qquad (2)$$

where $g\left(x_{\mathrm{train},i}\right)$ is the CNN output for the i-th training input, $t_{\mathrm{train},i}$ is the manually labeled pixel coordinate of the corner position, T is the set of indices for the training data, $\{a_i\}_{i \in A}$ is the set elements of which are variables a_i associated with i in the set A, $\mathrm{Med}(A)$ is the median of elements in the set A, and c_m is a sensitivity coefficient with respect to diagonal length determined solely by the magnification factor m. Importantly, c_m is independent of the content of the input image. Consequently, in this study, the aleatoric uncertainty is treated as a function of magnification only and does not vary with other input features.

The standard uncertainty in the diagonal length prediction, denoted by $u_{\mathrm{CNN}}(x)$, is calculated as the square root of the sum of the squares of the standard epistemic and aleatoric uncertainties:

$$u_{\mathrm{CNN}}(x) = \sqrt{u_{\mathrm{eps}}^2(x) + u_{\mathrm{ale}}^2(x)} \qquad (3)$$

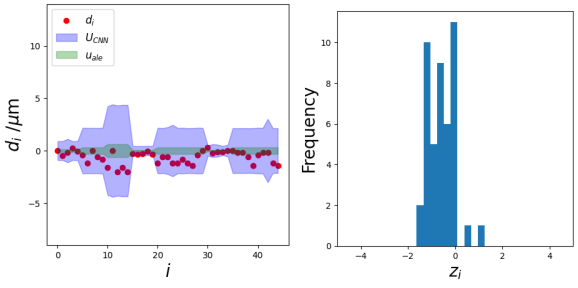
For simplicity, we do not handle other uncertainty sources associated with hardness measurement, as they are beyond the scope of this study. The expanded uncertainty is computed using a coverage factor $k = 1.96$, giving: $U_{\mathrm{CNN}}(x) = k u_{\mathrm{CNN}}(x)$

IV. RESULTS AND DISCUSSION

Two types of test data are evaluated: one measured on materials that are included in the training dataset, and another measured on materials not included in the training data. The results measured on the same materials as those in the training data are shown in Fig. 2. The data were obtained using images of reference blocks with nominal values of 200 HV, 600 HV, and 900 HV. Fig. 2(a) shows the relationship between the difference d_i between the measured diagonal lengths $y_{\mathrm{test},i}$ and predicted diagonal lengths $f\left(x_{\mathrm{test},i}\right)$ and the evaluated expanded uncertainty $U_{\mathrm{CNN}}(x_{\mathrm{test},i})$. The points represent the differences $d_i = f\left(x_{\mathrm{test},i}\right) - y_{\mathrm{test},i}$, where the blue shaded area represents the expanded uncertainty $U_{\mathrm{CNN}}(x_{\mathrm{test},i})$, and the green shaded area represents the standard aleatoric uncertainty $u_{\mathrm{ale}}(x)$. In comparison with the sensitivity coefficient c_m depending on magnification, the smallest standard epistemic uncertainty was $u_{\mathrm{eps}}\left(x_{\mathrm{test},11}\right)/c_m = 2.66$, and the largest was $u_{\mathrm{eps}}\left(x_{\mathrm{test},32}\right)/c_m = 3.88$.

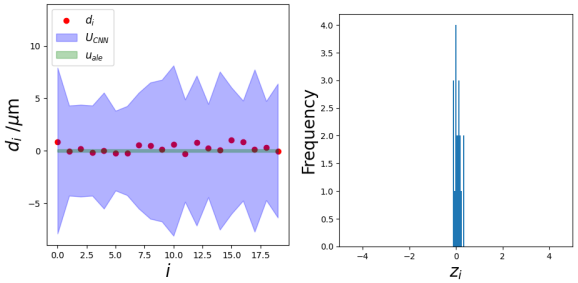
Since the standard aleatoric uncertainty was $u_{\mathrm{ale}}\left(x_{\mathrm{test},11}\right)/c_m = u_{\mathrm{ale}}\left(x_{\mathrm{test},32}\right)/c_m = 0.53$, the difference in standard epistemic uncertainty is significantly larger than the standard aleatoric uncertainty. This implies that, if a common epistemic uncertainty were quantified from the test data, this difference would not be reflected in the quantified uncertainty, and the uncertainty would be over- or under-evaluated. Therefore, it is necessary to evaluate the uncertainty according to the characteristics of each image, as done in this study.

To verify the evaluated uncertainty, Fig. 2(b) shows a histogram of the ratio between the difference d_i and the evaluated standard uncertainty $u_{\mathrm{CNN}}(x_{test,i})$, defined as $z_i = d_i / u_{\mathrm{CNN}}(x_{\mathrm{test},i})$.



(a) (b)

Fig. 2 Measurement of 200 HV, 600 HV and 900 HV reference blocks (a) relationship between d_i *and* $U_{CNN}(x_{test,i})$ *(b) histogram of* $z_i = d_i/u_{CNN}(x_{test,i})$



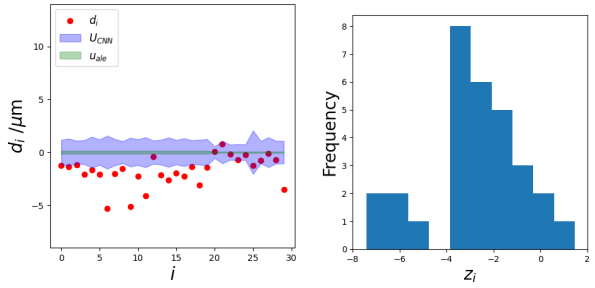
(a) (b)

Fig. 3 Measurement of CaF_2 *(a) relationship between* d_i *and* $U_{CNN}(x_{test,i})$ *(b) histogram of* $z_i = d_i/u_{CNN}(x_{test,i})$

For the case measured on the same material as the training data, all differences d_i fall within the expanded uncertainty $U_{\mathrm{CNN}}(x_{\mathrm{test},i})$. Furthermore, as seen in the histogram, the uncertainty is not excessively large. Therefore, it can be said that the estimated uncertainty is reasonable. It is considered that MCBN was applicable in this case.

Next, Figs. 3 and 4 show the results for two materials different from those in the training data: CaF_2 and Si_3N_4, respectively. The structure of Figs. 3 and 4 is the same as that of Fig. 2. As shown in Fig. 3(b), the maximum absolute value of z_i is less than 0.4. Therefore, in the case of CaF_2, it is presumed that either the aleatoric uncertainty, the epistemic uncertainty, or both were over-evaluated. On the other hand, in the case of Si_3N_4, the results indicate that the uncertainty was under-evaluated. There is a tendency for the predicted average diagonal length $f(x_{\mathrm{test},i})$ to show smaller values. Such bias should be evaluated as epistemic uncertainty.

Thus, the evaluation method in this study may not provide valid uncertainty evaluation for materials different from those used in training. Two possible reasons are considered. One is that the mini-batches used in the MCBN for evaluating epistemic uncertainty consist of training data (standard reference blocks and TiO_2). Since the input for materials different from the training data comes from a data region different from that of the training data, it is considered that the BNN using the MCBN did not perform properly. The second reason is that the aleatoric



(a) (b)

Fig. 4 Measurement of Si3N4 (a) relationship between d_i *and* $U_{CNN}(x_{test,i})$ *(b) histogram of* $z_i = d_i/u_{CNN}(x_{test,i})$

uncertainty is evaluated using the training data. Since the variation in measurement values differs depending on the material, aleatoric uncertainty should be evaluated for each material.

V. CONCLUSION

In this study, uncertainty evaluations in the Vickers hardness measurement using a convolutional neural network were conducted. To evaluate the uncertainty of the CNN outputs, it is necessary to evaluate two types of uncertainties: aleatoric uncertainty and epistemic uncertainty. In the present study, epistemic uncertainty was evaluated using Monte Carlo batch normalization, and aleatoric uncertainty was estimated from the difference between the training data and the predicted values.

For data obtained from the same materials as those used in the training dataset, it was confirmed that the proposed evaluation method allows for a valid assessment of uncertainty. The quantified standard uncertainty varied significantly for each image, indicating the necessity of evaluating measurement uncertainty not comprehensively, but individually for each image. On the other hand, for data obtained from materials different from those in the training dataset, a valid uncertainty assessment remains challenging, but it is an important future task.

REFERENCES

[1] Tanaka, Yukimi, Yutaka Seino, and Koichiro Hattori. "Automated Vickers hardness measurement using convolutional neural networks." The International Journal of Advanced Manufacturing Technology 109.5 (2020): 1345-1355.

[2] Thompson, Andrew, et al. "Uncertainty evaluation for machine learning." (2021).

[3] Hüllermeier, Eyke, and Willem Waegeman. "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods." Machine learning 110.3 (2021): 457-506.

[4] Teye, Mattias, Hossein Azizpour, and Kevin Smith. "Bayesian uncertainty estimation for batch normalized deep networks." International conference on machine learning. PMLR, 2018.

[5] The Linux Foundation, PyTorch, https://pytorch.org/ (accessed on 13th Mar. 2025).