Machine Learning-Based Measurement Models: A Novel Methodology for GUM-Based Uncertainty Evaluation

Leopoldo Angrisani, *Fellow, IEEE*, Pasquale Arpaia, *Senior Member, IEEE*, Sabatina Criscuolo, *Graduate Student Member, IEEE*, Mauro D'Arco, *Senior Member, IEEE*, Egidio De Benedetto, *Senior Member, IEEE*, Luigi Duraccio, *Member, IEEE*, and Annarita Tedesco

***Abstract*—The continuous advancement of computational technologies has accelerated the use of data-driven methods across various domains, including metrology. Machine Learning (ML) models, in particular, have demonstrated significant potential to enhance measurement processes, establishing themselves as fully fledged elements of *modern* metrology. However, in such ML-based measurements, the evaluation of measurement uncertainty requires additional considerations compared with the *conventional* approaches to measurement. The most widely adopted framework in metrology, i.e., the guide to the expression of uncertainty in measurement (GUM), is based on the assumption that the measurand must be defined in such a way that definitional uncertainty is negligible compared to the other components of measurement uncertainty. In conventional practice, this is achieved by selecting an appropriate relationship between the measurand and a set of input quantities, referred to as the *measurement model*. In contrast, in ML-based measurements, this relationship is not explicitly chosen by the practitioner but rather inferred from data. Consequently, should adherence to the GUM framework be intended, it is necessary to verify the validity of the GUM assumption and take corrective actions where appropriate. Based on these considerations, this article proposes a novel methodology that enables GUM-based uncertainty evaluation in ML-based measurements. The primary focus is on ML deterministic regression models, whose structure can be readily assimilated into measurement operations. The proposed methodology is applied to two representative case studies. The first case study involves an electric circuit for resistance measurement, where a comparison between ML-based approaches and a consolidated measurement model is provided to ensure consistency. The second case study concerns the estimation of the output power of a power plant, where no conventional measurement model is available. Overall, the proposed methodology enables proper integration of ML practices into the GUM framework, thus broadening the application domain of measurement processes.**

***Index Terms*—Artificial intelligence (AI), gum, machine learning (ML), measurement uncertainty, metrology, regression, uncertainty.**

I. Introduction

THE emergence of artificial intelligence (AI), and in particular machine learning (ML), has introduced new opportunities across various scientific disciplines [1]. Metrology is not exempt from this trend [2]. It has increasingly benefited from the application of ML-based methods as a means to measure, through indirect methods, physical quantities that are otherwise difficult or even impractical to measure directly [3]. Such situations typically arise when the measurand cannot be expressed in terms of a known functional relationship with directly measurable input quantities or when such a relationship exists but does not satisfy the metrological requirements imposed by the specific application context [4].

It should be noted, however, that the use of such ML-based methods for measurement applications raises significant considerations whenever there is a need to adhere to established metrological standards. This article specifically refers to the Guide to the Expression of Uncertainty in Measurements (GUM), which represents the most widely adopted framework in metrology.[1] According to the GUM, the evaluation of measurement uncertainty can be performed under the following assumption. The measurand must be defined such that the *definitional uncertainty* can be considered negligible in comparison with the other components of measurement uncertainty [5]. The GUM suggests defining the measurand through the establishment of an appropriate *measurement model*, namely a mathematical relationship among all quantities known to be involved in a measurement. In conventional metrology, measurement models are consciously selected by practitioners and are typically expressed in the form of closed-form analytical relationships. However, the GUM does not preclude the use of alternative model formulations, including

Received 29 September 2025; revised 5 November 2025; accepted 11 November 2025. Date of publication 25 December 2025; date of current version 2 January 2026. This work was supported by Italian Ministry of University and Research (MUR) through the Project "Made in Italy Circolare e Sostenibile," PNRR PE11, under Grant CUP E63C22002130007. The Associate Editor coordinating the review process was Dr. Jing Lei. *(Corresponding author: Egidio De Benedetto.)*

Leopoldo Angrisani, Pasquale Arpaia, Mauro D'Arco, Egidio De Benedetto, and Luigi Duraccio are with the Department of Information Technology and Electrical Engineering (DIETI), University of Naples Federico II, 80125 Naples, Italy (e-mail: egidio.debenedetto@unina.it).

Sabatina Criscuolo is with the Institute of Intelligent Industrial Technologies and Systems for Advanced Manufacturing, National Research Council of Italy, 20133 Milan, Italy, and also with the Department of Information Technology and Electrical Engineering (DIETI), University of Naples Federico II, 80125 Naples, Italy.

Annarita Tedesco is with the Department of Public Health, University of Naples Federico II, 80131 Naples, Italy.

Digital Object Identifier 10.1109/TIM.2025.3643048

[1] A different and less widely adopted framework is provided by the International Electrotechnical Commission (IEC). The considerations presented in this article may not fully apply to the IEC framework.

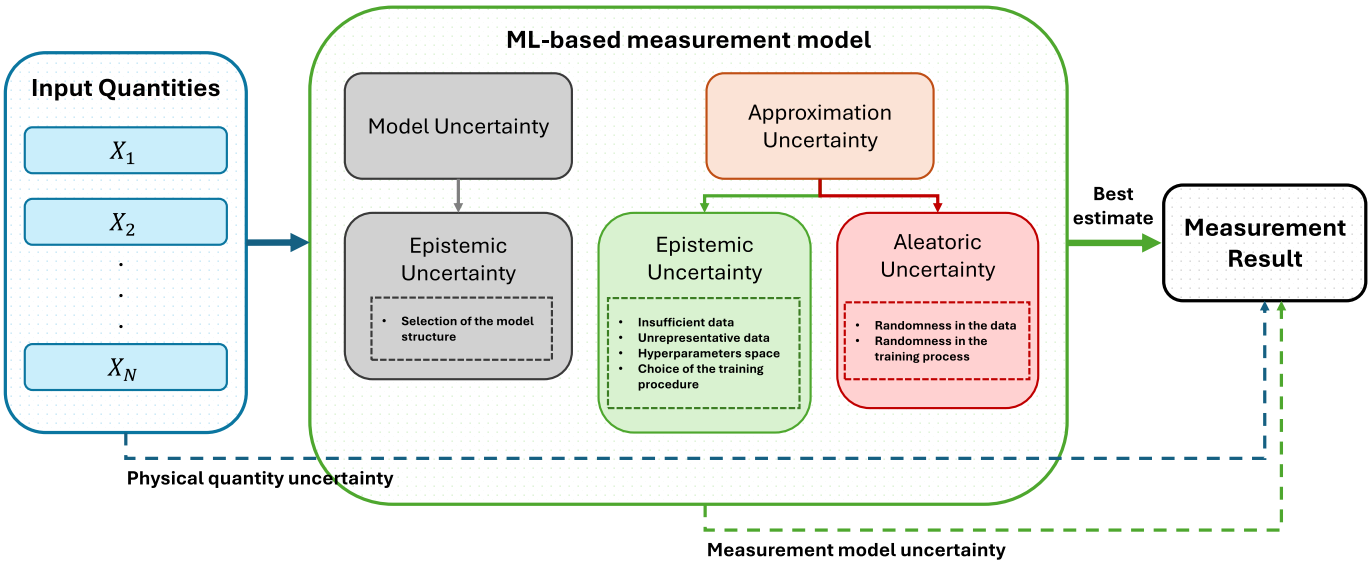

Fig. 1. Overview of the measurement uncertainty components in ML-based measurements. While the physical quantity uncertainty represents the conventional measurement uncertainty, the *measurement model uncertainty* constitutes an additional contribution, accounting for model and approximation uncertainties, which in turn consist of epistemic and/or aleatoric contributions. Both the physical quantity uncertainty and the measurement model uncertainty, alongside the best estimate of the measurand, constitute the measurement result.

those derived from empirical or data-driven approaches, provided that the assumption of negligible definitional uncertainty remains valid [6]. In this regard, as discussed in the relevant literature [7], an ML model holds the potential for being considered a valid measurement model. Nevertheless, unlike conventional measurement models, it is crucial to consider that such *ML-based measurement models* are not explicitly chosen by practitioners. Instead, they are inferred through a learning (also known as training) process that depends on multiple factors, including the amount and quality of the data provided [8]. It follows that, in the context of ML-based measurements, assessing whether the uncertainty related to ML-based models can be considered negligible (thus satisfying the GUM assumption) becomes challenging.

As a matter of fact, the uncertainty related to ML-based measurement models is multifaceted and has been categorized in different ways. A common taxonomy distinguishes between *model uncertainty*, which refers to the uncertainty associated with selecting the appropriate model structure, i.e., the choice of the mathematical function to be fit to the given training dataset, and *approximation uncertainty*, which refers to the different learning and approximation steps chosen during the training process [9], [10]. Another prevalent classification differentiates between *epistemic uncertainty* and *aleatoric uncertainty*. *Epistemic uncertainty* denotes the uncertainty stemming from incomplete knowledge about the underlying data-generating process or model structure. It is inherently reducible through the acquisition of additional data, refinement of the model, or an enhanced training process and domain understanding. *Aleatoric uncertainty*, on the other hand, arises from intrinsic variability or randomness in the observed phenomena. It reflects the intrinsic noise present in the data and, eventually, in the training process and is generally considered irreducible, even with the acquisition of additional information [11], [12]. These taxonomies are not mutually exclusive. As reported in [9], *model uncertainty* exhibits an *epistemic* contribution (arising from the selection of the model to be employed), whereas *approximation uncertainty* comprises both *epistemic* (e.g., arising from lack of sufficient training data, or selection of the the hyperparameters' space and training processes) and *aleatoric* contributions (e.g., arising from randomness in the data and training process, such as initialization of the model parameters and dropout strategies).

Hence, the uncertainty related to the ML-based measurement model, which can be named *measurement model uncertainty*, needs to be considered as an additional component in the overall measurement uncertainty budget, alongside the conventional one arising from the measurement of the input quantities, which can be named *physical quantity uncertainty*. These two components are illustrated in Fig. 1. As shown, the *measurement model uncertainty* comprises contributions from model and approximation uncertainties, which in turn consist of *epistemic* and/or *aleatoric* contributions. Therefore, the measurement result comprises the best estimate of the measurand together with its associated uncertainty, which encompasses both *physical quantity uncertainty* and *measurement model uncertainty* components, where the GUM explicitly assumes that the latter contribution is negligible compared with the former.

Building on these considerations, this article proposes a novel methodology enabling the application of GUM guidelines for uncertainty evaluation in ML-based measurements. Attention is dedicated to: 1) determining whether the measurement model uncertainty can be considered negligible with respect to the physical quantity uncertainty, in accordance with the GUM assumption; and 2) proposing corrective actions in cases where this condition is not met. This initial study focuses on deterministic regression models, whose mathematical

structure can be meaningfully interpreted within the framework of measurement operations [13]. Two representative case studies were considered to show the application of the methodology. The first case involves an electrical circuit used for resistance measurement, in which the performance of an ML-based measurement model is compared with that of an established measurement model to ensure consistency. The second case addresses the estimation of the output power of a power plant, where no conventional measurement model is available.

This article is organized as follows. Section II provides a background on ML-based regression models and related uncertainty sources. Section III describes the proposed methodology to evaluate the *measurement model uncertainty*. In Sections IV and V, the two aforementioned case studies are shown, along with the obtained experimental results. Finally, conclusions are drawn and future work is outlined.

II. Background

Regression analysis is a statistical technique for investigating and modeling the relationship between variables [14]. In the context of indirect measurement, regression analysis is particularly useful for measuring an unknown physical quantity Y, corresponding to the measurand, from a set of input physical quantities $(X_1, X_2, \ldots, X_N)$, which are typically more accessible or practical to measure. From a mathematical standpoint, the resulting functional relationship between Y and X corresponds precisely to the measurement model as defined in the GUM framework. According to this, the measurement model may be constructed using data-driven modeling techniques, and ML is arguably the most commonly used [15]. Thereby, this section provides a detailed overview of ML-based regression models, along with an in-depth discussion of the typical sources of uncertainty associated with their use.

A. ML-Based Regressors

ML regression is a supervised learning approach aimed at identifying a functional relationship that maps the space X, consisting of the *predictors* or *features* (i.e., the input physical quantities) to the output space Y, which contains continuous *target* values (i.e., the measurand). To achieve this, a *hypothesis space* H is defined, containing hypothesis functions $h : X \rightarrow Y$. The objective is to find the hypothesis $h^* \in H$ that best approximates the *true* underlying function f [9]. Thus, the process of estimating f, known as *training*, involves selecting a *suitable* estimator h^* by minimizing a loss function, commonly the mean squared error (mse), that quantifies the dispersion between predicted and observed values [16]. Once trained, the model can be used to predict outputs for new input instances, assuming these inputs are drawn from a distribution similar to that of the training data.

Based on the provided output at prediction time, ML-based regression models can be categorized into *deterministic* and *stochastic* [11], [15]. Deterministic models, such as simple feedforward artificial neural networks (ANNs) or support vector machines, are characterized by a fixed mapping from input to output. This means that for a given input, the output is always the same, with no variability between predictions for repeated estimations. On the other hand, stochastic models, including Bayesian neural networks and ensemble methods, introduce inherent variability in their predictions, generating probability distributions over the outcomes.

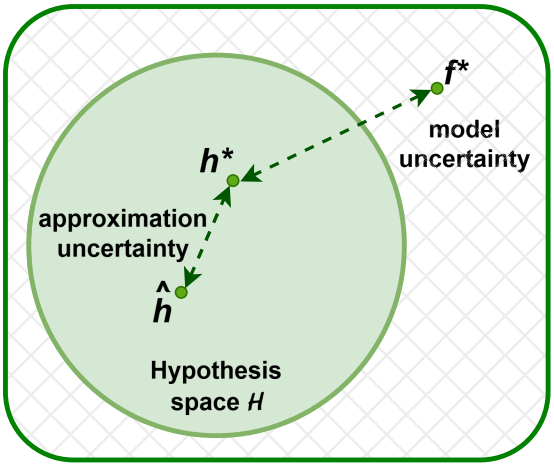

Fig. 2. Conventional representation of model and approximation uncertainties: the function f^* is the *ground truth*, $h^* \in H$ is the *best possible* learned function within the hypothesis space, and $\hat{h}$ is the *induced predictor* produced by the ML model [9].

B. Uncertainty Sources in ML-Based Deterministic Regressors

Focusing on ML-based deterministic regressors, it is highly desirable to evaluate the uncertainty associated with the prediction y_p for a given query instance $(x_{1p}, x_{2p}, \ldots, x_{Np}) \in X$. This uncertainty, which encompasses all sources of doubt affecting the predicted outcome,[2] is generally decomposed into *model uncertainty* and *approximation uncertainty* [9], [10].

Model uncertainty, which exhibits an *epistemic* uncertainty contribution [9], arises from the choice of the appropriate model class, specifically the selection of the hypothesis space H and, consequently, the mathematical form used to map inputs X to outputs Y. *Model uncertainty* captures the discrepancy between the learned function h^*, often referred to as the *best possible*, and the optimal function f^*, referred to as the *ground truth*, which best approximates the true underlying relationship f [9]. Thus, it could be stated that

$$h^* = f^* + \epsilon \tag{1}$$

with ϵ accounting for such discrepancy from the ground truth [9]. However, it is typically assumed that the hypothesis space H is correctly specified such that $f^* \in H$. Consequently, *model uncertainty* ϵ is neglected. As a matter of fact, this type of uncertainty is inherently challenging to quantify, as it involves expressing doubt over which hypothesis space might be the right one. In practice, conducting such an analysis is considered unfeasible [9].

[2]Unlike stochastic regressors, uncertainty evaluation in deterministic regressors is more challenging, since state-of-the-art methods such as variational inference, Monte Carlo dropout, or deep ensembles are not applicable [11].

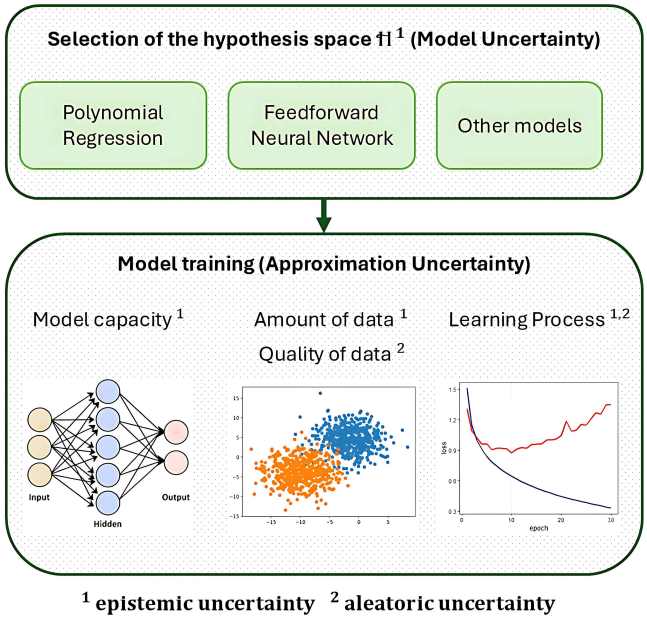

Fig. 3. Schematic of the relationship between model uncertainty, approximation uncertainty, epistemic uncertainty, and aleatoric uncertainty in ML-based measurement models, with an example of typical deterministic regressors (i.e., polynomial regression and feedforward neural network). Model uncertainty is typically neglected, whereas approximation uncertainty can be more readily quantified by addressing its epistemic and aleatoric contributions.

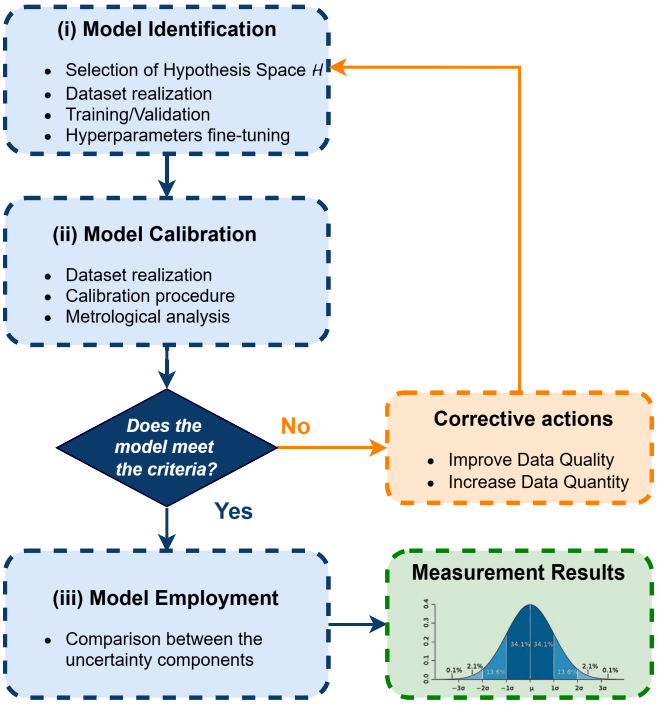

Fig. 4. Flowchart of proposed methodology for ML-based *measurement model uncertainty* evaluation.

On the other hand, the functional relationship h^* is inferred from the data through the training process. Due to various approximation and learning steps inherent in training, the resulting functional relationship $\hat{h}$, often referred to as the *induced predictor*, is itself only an estimate of h^*. Such discrepancy is known as *approximation uncertainty* and can be formalized by expressing the learned function as follows:

$$\hat{h} = h^* + e \tag{2}$$

where e accounts for the deviation from the best possible predictor [9]. In Fig. 2, a conventional graphical representation of model and approximation uncertainties is provided.

Notably, in the context of ML-based measurements, the learned function $\hat{h}$ corresponds precisely to the measurement model. Therefore, $\hat{h}$ can be referred to as the *ML-based measurement model*. However, since $\hat{h}$ is: 1) not chosen by the practitioner; 2) not known in closed form; and 3) strongly influenced by both the quality and quantity of the available training data, it is not possible to neglect such approximation uncertainty, which is composed, in turn, of both epistemic and aleatoric uncertainty contributions [9] Epistemic contribution arises from the lack of sufficient training data, the use of an insufficiently refined model in terms of selection of the hyperparameters, or the decisions made during the training process. Conversely, aleatoric contribution refers to data uncertainty arising from the inherent stochasticity of the observations and training process. This type of uncertainty typically cannot be reduced by increasing the amount of training data [9], [10]. However, the distinction between *aleatoric* and *epistemic uncertainty* is often ambiguous. As stated in [9], these uncertainty contributions should not be seen as absolute notions, but they depend on the given context that includes the model, feature set, data, and predictions. Variations in these contextual factors can alter the classification of uncertainty sources, with *aleatoric uncertainty* potentially manifesting as *epistemic uncertainty*, and vice versa. Overall, these considerations underscore the need for robust methodologies to quantify ML-based *measurement model uncertainty*, encompassing both *epistemic* and *aleatoric* contributions (summarized in Fig. 3) from a more holistic perspective.

III. Method

Fig. 4 illustrates the proposed methodology, which comprises three main steps. The first step is the identification of the *best* ML-based measurement model for the intended measurement task. Arguably, this is conducted following the established ML practices. The second step involves the evaluation of the uncertainty associated with the selected model, specifically through an appropriate model calibration. Conversely, this step is conducted according to metrological practices, since it is necessary to assess whether the ML-based measurement model satisfies the GUM assumption. Finally, the third step concerns the employment of the ML-based measurement model in the actual measurement process. Overall, this methodology integrates both ML and metrological practices in order to provide reliable measurement results, ensuring full compliance with the GUM framework.

A detailed description of each step is provided in Sections III-A–III-C.

A. Identification of the ML-Based Measurement Model

This first step comprises several tasks. Initially, it involves the selection of a suitable hypothesis space H, which defines

the class of models considered during training. The hypothesis space is typically determined by the chosen model architecture (e.g., the structure of a feedforward ANN or the kernel type in a support vector machine), as well as the range of hyperparameters to be optimized. As discussed in Section II, this stage primarily addresses *approximation uncertainty*, while *model uncertainty* is generally neglected under the assumption that the hypothesis space is sufficiently expressive to approximate the true underlying function [9]. Subsequently, in order to identify and train the most suitable ML-based measurement model for the specific measurement task, it is essential to construct a suitable training dataset that accurately reflects the conditions under which the ML-based measurement model is expected to operate. Guidelines for dataset construction are outlined as follows.

1) *High-Quality Instrumentation*: The dataset should be developed using accurate measuring systems. This ensures that the *aleatoric* contribution of the uncertainty remains within acceptable limits for the ML-based measurement model to be employed under real-world conditions.
2) *Sufficient Data Quantity*: The dataset should include a sufficiently large amount of data to ensure that the *epistemic* contribution of the uncertainty is reduced to acceptable levels, thereby supporting the deployment of the ML-based measurement model in practical scenarios.
3) *Representative Measurement Setup*: The dataset should accurately reflect the typical measurement setup, including relevant influence quantities under which the actual measurements will be performed.
4) *Range and Resolution*: The input quantities and corresponding measurand values should span the full expected operational range and resolution of the intended measurement task, ensuring adequate model generalization.

Once the dataset has been prepared, it should first be partitioned into distinct subsets for training and for validation. This partitioning is typically performed using techniques such as hold-out splitting or k-fold cross validation [16]. The training set is used to estimate the parameters of the ML-based measurement model (e.g., weights and biases in the case of an ANN), while the validation set is employed to assess the model's generalization capability and to guide model selection [16]. Following this step, the process of identifying and training the ML-based measurement model is undertaken. This involves hyperparameter optimization, commonly carried out using grid search or random search strategies [16], combined with validation procedures to evaluate predictive performance and to determine the most suitable model configuration. In the context of regression tasks, performance is typically quantified using metrics such as the root-mean-squared error (RMSE) and the mean absolute error (MAE), which express the deviation between predicted values and ground truth across the entire validation set. However, it is crucial to highlight that such performance metrics provide a *global* evaluation of the model's accuracy [10], without offering a direct quantification of the uncertainty associated with individual predictions. Consequently, while these metrics are useful during model development, they are not sufficient from a metrological standpoint, where each prediction is expected to be associated with an uncertainty value. Then, a more detailed uncertainty evaluation that takes into account not only the *physical quantity uncertainty* but also the *measurement model uncertainty* is required to ensure compliance with the GUM framework.

B. Calibration of the ML-Based Measurement Model

To assess the uncertainty related to the ML-based measurement model, this study proposes a procedure inspired to instrument calibration. However, given that an ML model plays the role of a measurement model, this process can be referred to as *model calibration*. As defined by the VIM [5], instrument calibration is an "operation that, under specified conditions, first establishes a relationship between quantity values with associated measurement uncertainties provided by measurement standards and corresponding indications with their respective measurement uncertainties. Then, it uses this information to establish a relation for obtaining a measurement result from an indication". The proposed *measurement model calibration* follows a similar structure. First, it establishes a relationship between the model predictions and the reference values of the measurand incorporating the associated uncertainties. Then, this relationship is used to evaluate the *measurement model uncertainty* and ensure its predictions can be reliably applied in metrological contexts.

Such a calibration process can be conducted, in principle, on the same validation set employed during model training. However, when a pretrained ML-based measurement model is provided without access to the original training/validation data, or when the available validation set does not satisfy the metrological conditions required for calibration (described in the following), it is necessary to adopt a broader perspective: calibration can be performed by an accredited calibration center, which is responsible for constructing a *calibration dataset* that mirrors the four characteristics of the training dataset mentioned in Section III-A. This ensures consistency with the operational conditions under which the model will be deployed.

The generation of the *calibration dataset* should be conducted with the following principle as a foundation; for each of the target values (i.e., the reference values of the measurand), the dataset should include a set of M distinct measurements (also referred to as *instances*) of the input quantities. This methodology allows for the evaluation of the mean of the ML-based measurement model predictions and the standard uncertainty at each target point. The result is analogous to a *calibration curve* for the measurement model, including a linear regression line that can be used to correct for any systematic effects during actual measurement operations. It is thus possible to assign an uncertainty value to each individual prediction of the ML model, which represents the uncertainty related to the ML-based measurement model, which has been named, for simplicity, *measurement model uncertainty*.

In the case that the metrological performance of the ML model does not meet predefined targets, either in terms of target uncertainty or deviation from reference values, the

procedure enables the implementation of corrective actions. These actions may involve: 1) improving data quality to reduce the aleatoric contribution of the uncertainty or 2) increasing the amount of data to reduce the epistemic contribution, or a combination of them. Thus, the calibration procedure is an iterative process between the model training and uncertainty evaluation steps. Corrections are implemented with the objective of achieving model performance that meets the GUM requirements.

C. Comparison of Uncertainty Components

Once the ML-based measurement model achieves satisfactory metrological performance, it can be deemed suitable for deployment within the intended application context. At this stage, conventional measuring systems typically account for uncertainty contributions arising from the measurement of the input quantities, which are propagated to the output using either the law of propagation of uncertainty [6] or Monte Carlo simulations [17].[3] However, in the case of ML-based measurement systems, the uncertainty associated with the measurement model itself also needs to be considered as an additional component in the overall uncertainty budget.

Consequently, as anticipated, the overall measurement uncertainty consists of two main components: 1) the physical quantity uncertainty, which encompasses the conventional contributions associated with the measurement of the input physical quantities; and 2) the measurement model uncertainty, which accounts for the uncertainty inherent in the ML-based measurement model itself. The latter can be evaluated starting from the best estimate of the measurand and the calibration diagram obtained through the model calibration procedure described in Section III-B.

By evaluating whether the *measurement model uncertainty* is negligible relative to the *physical quantity uncertainty*, it will be possible to determine the applicability of the GUM framework in its conventional form. However, if this condition is not met, additional procedures, which need to be properly investigated in future work, will be required to appropriately combine these two sources of uncertainty.

IV. Case Study #1

In this first case study, an electrical circuit suitable for resistance measurement is considered. The schematic of the circuit is shown in Fig. 5. The circuit includes: three resistances R_a, R_b, and R_c; a dc voltage source V_{CC}; a multimeter used to measure the voltage V_0 between nodes 1 and 2; and the unknown resistance R_X. This circuit is typically employed for resistance measurement following the Wheatstone bridge method. In such a configuration, the resistance R_c should be represented by a variable resistor, whose value is adjusted so that the measured voltage V_0 is as close as possible to 0 V, thereby maximizing the sensitivity of the measurement circuit. However, in this case study, such a balancing condition is not imposed. Instead, the relationship between the unknown resistance R_X and the other physical quantities R_a, R_b, R_c, V_{CC}, and V_0 is analyzed by solving the circuit in general form. Specifically, by applying Kirchhoff's laws, the following expression is derived:

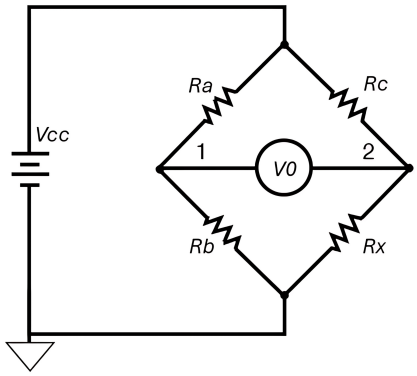

Fig. 5. Schematic of the electrical circuit considered in Case Study #1.

$$R_X = R_c \cdot \frac{\left(\frac{V_0}{V_{CC}} + \frac{R_b}{R_a+R_b}\right)}{1 - \left(\frac{V_0}{V_{CC}} + \frac{R_b}{R_a+R_b}\right)}. \tag{3}$$

This equation can be interpreted as a conventional measurement model, where the measurand R_X is a function of the input quantities R_a, R_b, R_c, V_{CC}, and V_0. The uncertainty associated with this model, due to nonidealities in Kirchhoff's laws, is reasonably negligible compared to the uncertainty associated with the measurement of the input quantities [18]. Therefore, in principle, there is no strict need to employ an ML-based measurement model. Nonetheless, the adoption of ML-based measurement models to estimate R_X from the input quantities can still serve as an illustrative scenario, in which the proposed methodology allows for a comparison between the *measurement model uncertainty* and *physical quantity uncertainty*. This enables the implementation of corrective actions to ensure that the GUM assumption is fulfilled.

In this case study, it is assumed that the goal is to perform resistance measurements within the range [100.0, 5000.0] Ω, with a target uncertainty of 1%. Consequently, it is expected: 1) that the uncertainty related to the measurement of the input quantities remains below this threshold and 2) that the uncertainty introduced by the ML-based measurement model is negligible in comparison. While the first condition can be satisfied by appropriately selecting the measurement setup, the second can be thoroughly investigated using the proposed methodology.

A. Training of the ML-Based Measurement Model

The dataset initially considered for training the ML-based measurement model consisted of resistance values of the measurand R_X ranging from 100.0 to 5000.0 Ω, in increments of 100.0 Ω, hence resulting in 50 distinct target values of R_X. The resolution with which R_X is provided is 0.1 Ω, which

[3]The law of propagation of uncertainty described in [6] allows for uncertainty evaluation when the measurement model is available in closed form, but it becomes impractical in complex or nonlinear cases. Supplement 1 to the GUM [17] recommends Monte Carlo simulations as a more general approach, particularly suitable when the functional form of the model is unknown. Since an ML-based measurement model is not available in closed form, Monte Carlo simulations represent the most effective strategy.

TABLE I
SEARCH SPACE ADOPTED HYPERPARAMETERS FINE-TUNING

Tuned Hyperparameters	Search Space
number of hidden layers	{1, 2, 3, 4}
number of neurons in each layer	{10, 20, 50, 100}
learning rate	$\{10^{-2}, 10^{-3}, 10^{-4}\}$
batch size	{32, 64, 100}

then constitutes the lower limit of the measurement uncertainty [6]. For each target value of R_X, the input quantities R_a, R_b, and R_c were randomly sampled from a uniform probability density function (pdf) within the range [100.0, 5000.0] Ω. Similarly, the dc voltage source V_{CC} was sampled from a uniform pdf over the interval [1.5, 5.5] V, to account for typical dc supply conditions. The voltage V_0 was computed using (3), with the addition of zero-mean Gaussian noise (standard deviation $\sigma = 1$ mV) to simulate nonidealities consistent with Kirchhoff's laws. This procedure yielded $M_t = 100$ distinct instances of the input quantities for each resistance value of the measurand R_X. Hence, the overall dataset comprises 5000 instances (i.e., 100 instances for each of the 50 resistance values of R_x).

After a preliminary recognition of the hypothesis space [9], the model selected for training was a feedforward ANN, chosen for its suitability in modeling complex, possibly nonlinear, input–output relationships that may arise in measurement systems [19]. To identify the optimal configuration of hyperparameters, a grid search procedure was performed along with a hold-out validation strategy, using an [80%, 20%] split between training and validation data. More specifically, Table I provides the tuned hyperparameters and the search spaces.

The activation function adopted in the hidden layers was the rectified linear unit (ReLU). MinMax normalization was employed to prevent potential scaling issues among the input quantities and to ensure numerical stability during the training process. Once the hyperparameter configuration yielding the best performance, i.e., the lowest mse between the predicted values and the target values, was identified, the ANN with such settings, trained on the training portion of the dataset, was identified as the ML-based measurement model.

In this case study, the optimal architecture was identified as an ANN consisting of four hidden layers with 100 neurons each, trained with a learning rate of 0.01 and a batch size of 32. This configuration achieved an mse of 6999.2 Ω^2.

However, as anticipated in Section II, the mse obtained cannot be considered a metric suitable for evaluating the uncertainty associated with the ML-based measurement model, as it is a global metric associated with the predictions made by an ML model across the entire dataset. Therefore, an appropriate model calibration procedure should be implemented to assess *measurement model uncertainty*.

B. Calibration of the ML-Based Measurement Model

The dataset employed for model calibration was developed under the same conditions as the training dataset, specifically with regard to the selection of values for the input quantities and the measurand R_X, ensuring consistency in terms of range, resolution, and PDFs. This design choice follows standard practice in ML, where both training and evaluation data are sampled from the same distribution to ensure a reliable assessment of model performance within its intended domain of applicability [16]. As for the dataset size, $M_c = 30$ instances of the input quantities were generated for each of the 50 target resistance values of the measurand R_X. Hence, for each of these 50 target values, 30 predictions were produced, each corresponding to a distinct instance of the input quantities. These predictions were then used to evaluate, for each resistance value of R_X: 1) the measurement bias, i.e., the difference between the reference value of R_X and the mean of the predictions, and 2) the standard uncertainty, i.e., the dispersion of the prediction mean, determined through a type-A evaluation. Starting from this information, it was possible to construct a fully fledged calibration diagram of the ML-based measurement model, from which a linear regression line was leveraged to correct for any systematic effects during actual measurement operations.

These results are reported in Fig. 6(a). As evident from the results, the evaluated standard uncertainties do not meet the target uncertainty requirements established for the case study. Specifically, for numerous mean predicted values of R_X, the corresponding standard uncertainty approached the 1.0% threshold. Therefore, appropriate corrective actions should be undertaken to ensure compliance with the prescribed metrological performance criteria.

C. Corrective Action

The chosen strategy for this case study was to act on the *epistemic* contribution of the *measurement model uncertainty*. Specifically, the training dataset was enriched by decreasing the step size between consecutive target values of the measurand R_X from 100.0 to 1.0 Ω. As a result, instead of 50 target values, a total of 4900 target values were produced. Consequently, by maintaining $M_t = 100$ instances for each target value, the overall dataset was expanded to 490 000 instances. The range, resolution, and PDFs associated with the input quantities remained consistent with those adopted in the previous step.

Following the same grid search procedure, with the search space defined in Table I, and employing the same hold-out validation strategy and normalization used in the initial training stage, a new ANN was identified and trained on the enriched dataset. The resulting optimal architecture consisted of four hidden layers with 50 neurons each, trained with a learning rate of 0.001 and a batch size of 64, achieving an mse of 304.1 Ω^2. Subsequently, model calibration was performed on the same calibration dataset in order to compare the metrological performance of the newly trained model. As illustrated in Fig. 6(b), the *measurement model uncertainty* was significantly reduced, thereby satisfying the prescribed uncertainty requirement.

D. Employment of the ML-Based Measurement Model and Uncertainty Evaluation

The proposed methodology enabled, during actual measurement operations, a comparison between the *physical quantity*

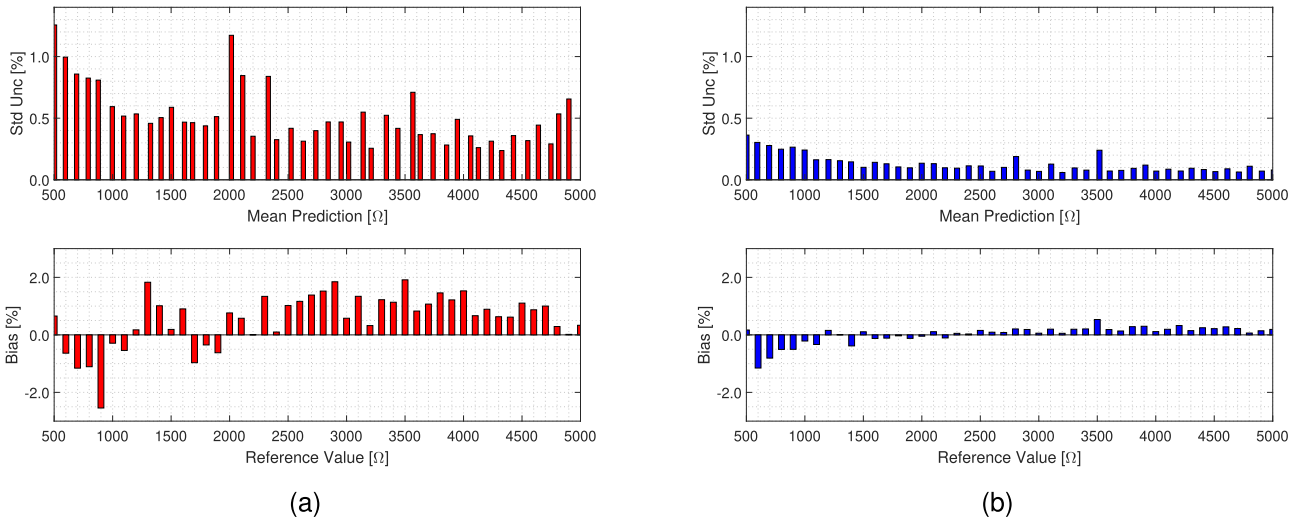

Fig. 6. Standard uncertainty and bias resulting from the model calibration of the ML-based measurement model #1 (a) and #2 (b) Case Study #1.

TABLE II
MEASUREMENTS OF THE INPUT QUANTITIES FOR THE EMPLOYMENT OF THE ML-BASED MEASUREMENT MODEL IN CASE STUDY #1

Input Quantity	Best Estimate	Standard Uncertainty	PDF
R_a	1502.0 Ω	8.7 Ω	Uniform
R_b	1503.0 Ω	8.7 Ω	Uniform
R_c	500.6 Ω	2.9 Ω	Uniform
V_{CC}	5.0420 V	$3{\cdot}10^{-4}$ V	Uniform
V_0	0.94905 V	$2{\cdot}10^{-5}$ V	Uniform

TABLE III
RESULTS OBTAINED THROUGH CONVENTIONAL AND ML-BASED MEASUREMENT MODELS IN CASE STUDY #1

Model	Best Estimate [Ω]	MCM Unc. [%] [a]	Model Unc. [%] [b]
Conventional Meas. Model	1106.0	0.57	n.a.
ML-based Meas. Model #1	1148.2	0.43	0.71
ML-based Meas. Model #2	1108.5	0.47	0.20

[a] MCM Unc. represents the *physical quantity uncertainty*, *i.e.*, the uncertainty propagated by using Monte Carlo Method (MCM) simulations. It is represented in the form of a relative standard uncertainty.
[b] Model Unc. [%]) represents the *measurement model uncertainty*, *i.e.*, the uncertainty associated to the ML-based measurement models. It is represented in the form of a relative standard uncertainty.

uncertainty and the *measurement model uncertainty*. For this case study, concerning the measurement of the unknown resistance R_X,[4] decade resistance boxes were employed for the quantities R_a, R_b, and R_c (1% tolerance), along with a dc power supply (Agilent 61100A) and a digital multimeter (Keysight 344401A). In Table II, the best estimates of the measured input quantities are reported, along with their associated standard uncertainties and the type of PDFs to be considered (derived from instrument specifications). In this way, the distributions of the input quantities were propagated through: 1) the conventional measurement model; 2) the ML-based measurement model #1, i.e., the model trained before dataset expansion; and 3) the ML-based measurement model #2, i.e., the model obtained after corrective actions. The propagation of the distributions was performed using Monte Carlo simulations, with a number of samples equal to 10^6. For each of the three resulting PDFs of the measurand, it was obtained: 1) the best estimate as the arithmetic mean (with correction of systematic effects); 2) the standard uncertainty as the standard deviation (representing the *physical quantity uncertainty*); and 3) the *measurement model uncertainty* (solely for the two ML-based measurement models). The experimental results obtained are reported in Table III. As expected, the best estimate provided by the ML-based measurement model #2 is closer to that of the conventional measurement model compared to ML-based measurement model #1. Furthermore, the *measurement model uncertainty* associated with ML-based measurement model #2 is significantly reduced with respect to model #1 and is also lower than the *physical quantity uncertainty*. Overall, the result obtained by ML-based measurement model #1 is not compatible with the conventional measurement model, whereas the one obtained by ML-based measurement model #2 exhibits compatibility. This confirms that the proposed methodology proved effective in enabling the ML-based measurement model to comply with the GUM framework.

[4]For the unknown resistance R_X, a resistance of nominal value of 1100 Ω was chosen as it is one of the possible target values included in both the training datasets.

V. CASE STUDY #2

The second case study is an example of the challenges in modern metrology when conventional measurement models are either unavailable or impractical. Specifically, it concerns the measurement of the full load electrical power output of a combined cycle power plant (CCPP) operating under base load conditions [20], as sketched in Fig. 7. Accurate measurements of the full load electrical power output are crucial for optimizing plant efficiency, maximizing economic returns from available energy, and ensuring reliability and sustainability of the system. In this scenario, a physically based measurement model would require solving a complex system of nonlinear thermodynamic equations.

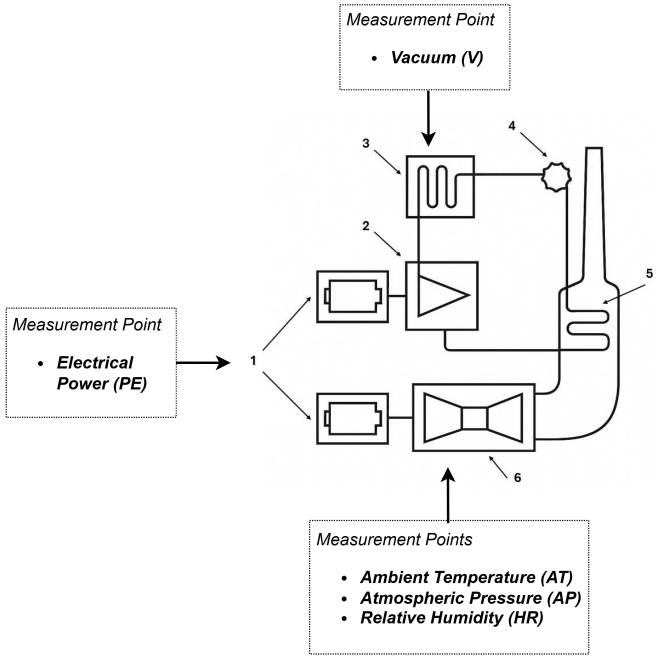

Fig. 7. Schematic of a generic CCPP in Case Study #2 (Legend: 1—electric generators, 2—steam turbine, 3—condenser, 4—pump, 5—boiler/heat exchanger, and 6—gas turbine).

The resulting computational burden and lack of tractability make the development of an explicit analytical model infeasible for practical purposes. This motivates the adoption of data-driven models and highlights the need to extend the GUM metrological framework to these approaches. Specifically, the gas turbine power output is particularly sensitive to ambient conditions, namely ambient temperature (AT), atmospheric pressure (AP), and relative humidity (HR), while the steam turbine output depends primarily on the vacuum pressure (V) in the condenser. Hence, these quantities can be considered as the input quantities for measuring the full load electrical power output by means of an ML-based measurement model. Also, in this case study, it is assumed that the goal is to perform measurements with a target uncertainty of 1%.

The dataset under analysis is publicly available at [21]. Regarding the input quantities, AT ranges from 1.81 °C to 37.11 °C, AP ranges from 992.89 to 1033.30 mbar, RH spans from 25.56% to 100.00%, and V ranges from 25.36 to 81.56 cm Hg. The measurand values range from a minimum of 420.26 MW to a maximum of 495.76 MW. In order to illustrate the applicability of the proposed methodology, approximately 70% of the original dataset was allocated to the identification and training phase of the ML-based measurement model, while the remaining portion was reserved for the model calibration procedure and on-field testing. Furthermore, zero-mean Gaussian noise was added to each value of the physical quantities in the dataset (relative standard deviation $\sigma_{\text{rel}} = 10\%$) to simulate the measurement uncertainty associated with data acquisition. The choice of Gaussian noise was motivated by the lack of detailed information on the instrumentation used to acquire the data. Subsequently, the measurand values were rounded to the nearest integer to ensure an adequate number of instances for each measurand value, specifically, $M_t = 100$ instances for each measurand value in the training dataset and $M_c = 30$ instances for each measurand value in the calibration dataset.

A. Training of the ML-Based Measurement Model

Also in this case study, the model selected for training was a feedforward ANN, with the ReLU employed as the activation function in the hidden layers. The same considerations adopted in case study #1 regarding hyperparameter optimization within the search space defined in Table I, data normalization, and the train–validation split were applied. The optimal ANN architecture initially identified consisted of four hidden layers with ten neurons each, trained with a learning rate of 0.01 and a batch size of 32, achieving an mse of 31.1 MW^2.

B. Calibration of the ML-Based Measurement Model

Once the ML-based measurement model was identified, it underwent the model calibration procedure, from which both the bias and the relative standard uncertainty were estimated for each target value of the measurand. These results are reported in Fig. 8(a). Regarding the measurement model uncertainty, its value was already significantly below the target uncertainty of 1%. However, the majority of the values predicted by the ML-based measurement model fell within the range [440, 470] MW, whereas the measurand values in the calibration dataset (used as a reference in the calibration procedure) were approximately distributed between 430 and 475 MW. In fact, the relative bias was considerably higher at the boundaries of the range compared to the intermediate values. Consequently, a corrective action was undertaken to mitigate this issue.

C. Corrective Action

In this case study, for illustrative purposes, the focus was placed on the *aleatoric* contribution of the *measurement model uncertainty*. Specifically, the training dataset was modified by reducing the relative standard deviation of the added Gaussian noise from 10% to 5% to simulate a practical condition involving the use of higher quality instrumentation. Hence, the training procedure was conducted under the same conditions as the initial training. The resulting ML-based measurement model was a feedforward ANN with two hidden layers and ten neurons per layer, trained with a learning rate of 0.01 and a batch size of 64, achieving an mse of 21.8 MW^2. Subsequently, the model underwent the calibration procedure using the same calibration dataset initially used. As shown in Fig. 8(b), the distribution of the predicted values improved, with the predicted range being closer to that of the reference values.

D. Employment of the ML-Based Measurement Model and Uncertainty Evaluation

Analogously to case study #1, the proposed methodology was exploited to compare the *physical quantity uncertainty*

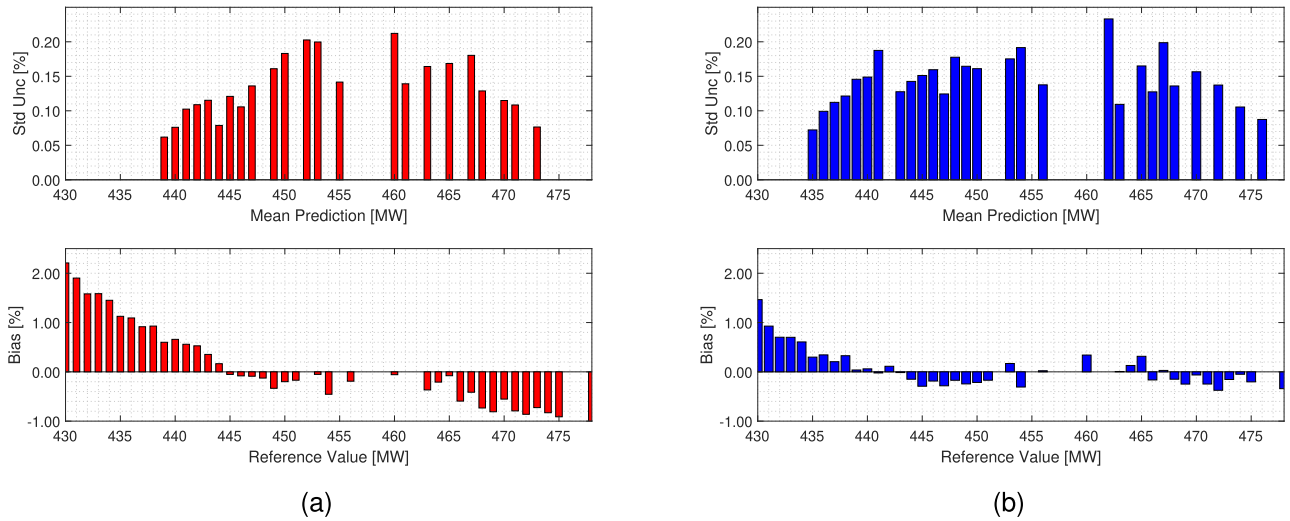

Fig. 8. Standard uncertainty and bias resulting from the model calibration of the ML-based measurement model #1 (a) and #2 (b) Case Study #2.

TABLE IV
MEASUREMENTS OF THE INPUT QUANTITIES FOR THE EMPLOYMENT OF THE ML-BASED MEASUREMENT MODEL IN CASE STUDY #2

Input Quantity	Best Estimate	Standard Uncertainty	PDF
AT	27.17 °C	0.29 °C	Uniform
AP	1011.4 mbar	0.2 mbar	Uniform
RH	65.32 %	0.29 %	Uniform
V	61.02 cm Hg	0.04 cm Hg	Uniform

TABLE V
RESULTS OBTAINED THROUGH ML-BASED MEASUREMENT MODELS IN CASE STUDY #2

Model	Best Estimate [MW]	MCM Unc. [%] [a]	Model Unc. [%] [b]
Reference	453	n.a.	n.a.
ML-based Meas. Model #1	450	0.80	0.14
ML-based Meas. Model #2	451	0.77	0.14

[a] MCM Unc. represents the *physical quantity uncertainty*, *i.e.*, the uncertainty propagated by using Monte Carlo Method (MCM) simulations. It is represented in the form of a relative standard uncertainty.
[b] Model Unc. [%]) represents the *measurement model uncertainty*, *i.e.*, the uncertainty associated to the ML-based measurement models. It is represented in the form of a relative standard uncertainty.

with the *measurement model uncertainty* during the employment stage of the ML-based measurement model. In this case study, concerning the measurement of the output power PE, it was assumed, for illustrative purposes, that the temperature quantity AT was measured using a *STTS22H* sensor, the relative humidity RH using a HDC3120 sensor, the vacuum level V using a DVR2pro vacuum gauge, and the AP using a LPS22DF sensor. Table IV reports the best estimates of the measured input quantities, along with their associated standard uncertainties derived from the instrument specifications and the assumed PDFs. The distributions of the physical input quantities were propagated through both ML-based measurement models, i.e., the models trained before and after the corrective actions, by means of Monte Carlo simulations with 10^6 samples. As in case study #1, the best estimate was obtained (with a correction of systematic effects), along with the *physical quantity uncertainty* and the *measurement model uncertainty*. The experimental results are presented in Table V. As shown, the best estimate produced by the ML-based measurement model #2 is closer to the reference value, while the total uncertainty contributions for both models are comparable. In any case, it is confirmed that the uncertainty attributable to the ML-based measurement model (0.14%) can be regarded as negligible[5] with respect to the uncertainties associated with the measurement of the input quantities (0.77%), thus fulfilling the conditions prescribed by the GUM framework.

VI. CONCLUSION AND FUTURE WORK

This article proposed a novel methodology enabling the application of GUM guidelines for uncertainty evaluation in ML-based measurements. Unlike conventional measurement models, whose definitional uncertainty is typically negligible by construction, ML-based measurement models infer the relationship between the measurand and the related physical input quantities from data. This fundamental difference necessitates: 1) thorough evaluation of the uncertainty associated with such ML-based measurement models (defined as measurement model uncertainty) and 2) comparison with the conventional physical quantity uncertainty in order to verify whether the GUM assumption is met. Focusing on deterministic regression models, the efficacy of the approach was demonstrated through two case studies: 1) an electrical circuit used for resistance measurements and 2) the measurement of the full load electrical power output of a CCPP. These case studies showed the capability of the methodology to validate ML models within the GUM framework and to identify when corrective actions are required. Specifically, the first case study confirmed that ML models can achieve consistency with established analytical measurement models, ensuring negligible model-related uncertainty. The second case study, in turn, highlighted the practical necessity of data-driven measurement

[5] As outlined in the GUM [6], an uncertainty component may be ignored if its contribution to the combined standard uncertainty of the measurement result is insignificant.

models when explicit physical models are infeasible, providing a path to comply with metrological standards despite the absence of conventional models. In both case studies, the ML model applied was a feedforward ANN. Future work will explore: 1) the application of the methodology with other ML architectures, such as recurrent neural networks, to better demonstrate its generality and 2) how to extend the proposed methodology to incorporate stochastic regressors as ML-based measurement models. Additionally, future studies will investigate how to combine the two uncertainty contributions when the assumptions of the GUM framework are infeasible to satisfy.

REFERENCES

[1] R. I. Mukhamediev et al., "Review of artificial intelligence and machine learning technologies: Classification, restrictions, opportunities and challenges," *Mathematics*, vol. 10, no. 15, p. 2552, Jul. 2022.

[2] M. Khanafer and S. Shirmohammadi, "Applied AI in instrumentation and measurement: The deep learning revolution," *IEEE Instrum. Meas. Mag.*, vol. 23, no. 6, pp. 10–17, Sep. 2020.

[3] A. Rashed and S. Shirmohammadi, "A novel method to estimate measurement error in AI-assisted measurements," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (I2MTC)*, May 2022, pp. 1–5.

[4] S. Shirmohammadi, "Measurement methodology: Visualizing uncertainty in machine learning-assisted measurements," *IEEE Instrum. Meas. Mag.*, vol. 26, no. 7, pp. 20–27, Oct. 2023.

[5] Joint Committee on Guides in Metrology (JCGM). (2012). *International Vocabulary of Metrology—Basic and General Concepts and Associated Terms (VIM)*. [Online]. Available: https://www.bipm.org/en/publications/guides/vim.html

[6] Joint Committee for Guides in Metrology (JCGM). (2008). *Evaluation of Measurement Data—Guide To the Expression of Uncertainty in Measurement (GUM)*. [Online]. Available: https://www.bipm.org/documents/20126/2071204/JCGM100 02008E.pdf

[7] T. Adel, S. Bilson, M. Levene, and A. Thompson, "Trustworthy artificial intelligence in the context of metrology," in *Producing Artificial Intelligent Systems: The Roles of Benchmarking, Standardisation and Certification*. Cham, Springer, 2024, pp. 53–75.

[8] T. M. Mitchell and T. M. Mitchell, *Machine Learning*, vol. 1. New York, NY, USA: McGraw-Hill, 1997.

[9] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Mach. Learn.*, vol. 110, no. 3, pp. 457–506, Mar. 2021.

[10] E. Buchicchio, P. Carbone, A. De Angelis, F. Santoni, and A. Moschitta, "Uncertainty quantification in AI-based measurement systems," *IEEE Instrum. Meas. Mag.*, vol. 28, no. 3, pp. 52–59, May 2025.

[11] J. Gawlikowski et al., "A survey of uncertainty in deep neural networks," *Artif. Intell. Rev.*, vol. 56, no. 1, pp. 1513–1589, 2023.

[12] A. Thompson et al., "Uncertainty evaluation for machine learning," Nat. Phys. Lab. (NPL), Teddington, U.K., Tech. Rep. MS 34, 2021.

[13] A. Thompson, "Analytical results for uncertainty propagation through trained machine learning regression models," *Measurement*, vol. 234, Jul. 2024, Art. no. 114841.

[14] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*. Hoboken, NJ, USA: Wiley, 2021.

[15] S. Shirmohammadi and H. Al Osman, "Machine learning in measurement part 1: Error contribution and terminology confusion," *IEEE Instrum. Meas. Mag.*, vol. 24, no. 2, pp. 84–92, Apr. 2021.

[16] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*, vol. 4. Cham, Switzerland: Springer, 2006.

[17] Joint Committee for Guides in Metrology (JCGM). (2008). *Evaluation of Measurement Data—Supplement 1 To the 'Guide To the Expression of Uncertainty in Measurement'—Propagation of Distributions Using a Monte Carlo Method (GUM)*. [Online]. Available: https://www.bipm.org/documents/20126/2071204/JCGM1012008E.pdf/325dcaad-c15a-407c-1105-8b7f322d651c

[18] R. S. Eisenberg, "Kirchhoff's law can be exact," 2019, *arXiv:1905.13574*.

[19] P. Daponte and D. Grimaldi, "Artificial neural networks in measurements," *Measurement*, vol. 23, no. 2, pp. 93–115, Mar. 1998.

[20] P. Tüfekci, "Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods," *Int. J. Electr. Power Energy Syst.*, vol. 60, pp. 126–140, Sep. 2014.

[21] *Combined Cycle Power Plant*, Dept. Comput. Eng., Fac. Çorlu Eng., Namık Kemal Univ., Tekirdağ, Turkey, 2014. [Online]. Available: https://archive.ics.uci.edu/dataset/294/combined+cycle+power+plant

Open Access funding provided by 'Università degli Studi di Napoli "Federico II"' within the CRUI CARE Agreement