Expert Systems With Applications 224 (2023) 120013

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

New metric formulas that include measurement errors in machine learning for natural sciences

Umberto Michelucci [a,c,*], Francesca Venturini [a,b]

[a] TOELT LLC, Birchlenstrasse 25, 8600 Dübendorf, Switzerland
[b] Institute of Applied Mathematics and Physics, Zürich University of Applied Sciences, Technikumstrasse 9, 8400 Winterthur, Switzerland
[c] Computer Science Department, Lucerne University of Applied Sciences and Arts, Werftestrasse 4, 6002 Lucerne, Switzerland

ARTICLE INFO

MSC:
68T01

Keywords:
Error propagation
Machine learning
Measurement errors
Regression
Classification
Computational physics

ABSTRACT

The application of machine learning to physics problems is widely found in the scientific literature. Both regression and classification problems are addressed by a large array of techniques that involve learning algorithms. Unfortunately, the measurement errors of the data used to train machine learning models are almost always neglected. This leads to estimations of the performance of the models (and thus their generalization power) that is too optimistic since it is always assumed that the target variables (what one wants to predict) are correct. In physics, this is a dramatic deficiency as it can lead to the belief that theories or patterns exist where, in reality, they do not. This paper addresses this deficiency by deriving formulas for commonly used metrics (both for regression and classification problems) that take into account measurement errors of target variables. The new formulas give an estimation of the metrics which is always more pessimistic than what is obtained with the classical ones, not taking into account measurement errors. The formulas given here are of general validity, completely model-independent, and can be applied without limitations. Thus, with statistical confidence, one can analyse the existence of relationships when dealing with measurements with errors of any kind. The formulas have wide applicability outside physics and can be used in all problems where measurement errors are relevant to the conclusions of studies.

1. Introduction

The main goal of training a supervised machine learning (ML) model is to find a relationship between a set of M inputs x_i (with $i = 1, \dots, M$) and some outputs y_i (with $i = 1, \dots, M$). The values of the variable to be predicted are often called labels in the literature. In general, x_i and y_i can be multidimensional; for simplicity, in this paper, the output variable is assumed to be a real number $y_i \in \mathbb{R}$. In any application of ML to a physics problem the output variables y_i will be either the direct result of a measurement or determined through some calculations from multiple ones. For example, y_i could be the oxygen concentration or temperature of a gas (Michelucci & Venturini, 2019), oxide glass-forming ability (Wilkinson et al., 2022), temperature in melt-pool fluid dynamics (Zhu, Liu, & Yan, 2021), or a measure of the dissolution kinetics of gases (Krishnan et al., 2018).

ML is used in physics in a large number of cases. For example, to locate phase transitions without any physical knowledge (Carrasquilla & Melko, 2017; Morningstar & Melko, 2018; Tanaka & Tomiya, 2017), to select events in collisions (Baldi, Bauer, Eng, Sadowski, & Whiteson, 2016; de Oliveira, Kagan, Mackey, Nachman, & Schwartzman, 2016) and to flavour tagging (Guest et al., 2016) in particle physics. In cosmology, ML has been applied, for example, to estimate the photometric redshift (Carrasco Kind & Brunner, 2013; Collister et al., 2007) and to predict fundamental cosmological parameters based on the dark matter spatial distribution (Ravanbakhsh et al., 2016). The list of examples is incredibly long, and applications can be found in almost all fields of physics. For an extensive review, the interested reader is referred to Carleo et al. (2019).

Apart from a few papers (Luo, Lorentzen, & Bhakta, 2021; Zhang, Xiao, Luo, & He, 2022), typically the performance of the models is reported without taking into account the measurement errors on the labels. Target variables that have errors present a certain uncertainty, and it is not clear how this uncertainty propagates to the metrics used to measure the performance of ML models. Research starts to indicate that in many physics problems, ignoring measurement errors may lead to an underestimation of ML model uncertainties (Ghosh & Nachman, 2022).

The problem is of fundamental relevance since typically models are trained on a specific dataset, typically split into a training and

[*] Corresponding author at: TOELT LLC, Birchlenstrasse 25, 8600 Dübendorf, Switzerland.
E-mail addresses: umberto.michelucci@toelt.ai (U. Michelucci), vent@zhaw.ch (F. Venturini).

https://doi.org/10.1016/j.eswa.2023.120013
Received 15 October 2022; Received in revised form 13 February 2023; Accepted 29 March 2023
Available online 1 April 2023

test part, and results are given in terms of specific metrics, such as the mean squared error (MSE), the mean absolute error (MAE) for regression problems or the accuracy (a) for classification problems (or slightly different metrics in case of unbalanced datasets of multi-class classification problems). The problem is that in almost all cases, measurement errors on the variables to be predicted (y_i) are ignored. This will lead to overly optimistic estimates of the mentioned metrics (such as the MSE, MAE or a).

The problem of noisy labels (a different kind of problem than the one analysed in this paper) is a relatively widely researched topic (Cour, Sapp, & Taskar, 2011; Menon, Rooyen, Ong, & Williamson, 2015; Natarajan, Dhillon, Ravikumar, & Tewari, 2013; Yao, Yang, Han, Niu, & Kwok, 2020; Zheng et al., 2020). Some articles deal with methods to identify wrongly labelled observation (Bahri, Jiang, & Gupta, 2020) and some try to define modified loss functions that can deal with noise (Liu & Tao, 2016). All efforts are directed towards understanding how to get better model performance or model stability when labels are noisy. However, none of these works addresses the problem of how labels errors can influence the metrics evaluated. Particularly in physics (but in all sciences for that matter), measurement errors must be included in any results as any physicist learns early in his or her career. This paper addresses this deficiency and provides formulas that take into account errors to better estimate the metrics most commonly used in ML, both for regression and classification problems.

The main contributions of this paper are four. Firstly, formulas are given for the MSE, the MAE, the accuracy a, and their variances that take into account measurement errors on the variable to be predicted y_i. These formulas are of general validity and are independent of the ML model used. Secondly, all the formulas are fully derived mathematically with a statistical approach. Thirdly, an a-priori mathematical derivation for the formulas is given in the appendices of the paper. Finally, guidelines are presented on how to use those formulas.

2. Problem formulation and notation

This paper considers the following thought experiment: find a relationship between a set of M inputs x_i (with $i = 1, \dots, M$) and some outputs y_i (with $i = 1, \dots, M$) that are assumed to be independent. The typical steps to achieve this are to split the dataset into training and test datasets, train the model on the training dataset, and validate the results by applying the model to the test dataset, namely, on unseen data. The training is done by minimizing an appropriate loss function, typically the MSE or MAE for regression, or the cross entropy for classification. Several metrics can be evaluated to assess the performance of the trained model. In this paper, the metrics most commonly used are discussed: MSE, MAE and accuracy a.

An important step to evaluate the generalization properties of the trained model is to perform a cross-validation (Michelucci & Venturini, 2021). An example of cross-validation is the split-train approach, achieved by performing the dataset split multiple times thus obtained multiple training and test datasets. Each time a new model is trained and its performance evaluated on the test dataset. Naturally, every time a new split is used, the model changes. Looking at the metrics obtained by evaluating them on the multiple test datasets, one can get an indication on the average performance of possible models obtained with the initial datasets. Unfortunately, this method does not take into account in any way the errors that are inevitably present on the y_i. Therefore, the metrics' values will lead to an overly optimistic impression of the goodness of the model.

Let us introduce a measurement error on the labels y_i with the random variable ϵ_i

$$y_i = \overline{y}_i + \epsilon_i \tag{1}$$

where $\overline{y}_i$ is the average value of y_i and ϵ_i follows a normal distribution with average zero and variance σ_i^2, or more formally,

$$\epsilon_i \sim \mathcal{N}(0, \sigma_i^2) \tag{2}$$

or in other words

$$y_i \sim \mathcal{N}(\overline{y}_i, \sigma_i^2). \tag{3}$$

This is based on the hypothesis that measurement errors follow a Gaussian distribution (Taylor, 1997). This work provides formulas for the expected value and variance of metrics, specifically MSE, MAE, and a, over the distribution of the random variable ϵ_i.

In this paper, two generic problems are considered: a regression problem and a classification problem. A generic **regression problem** has the objective of predicting a continuous variable using a set of M observations tuples (x_i, y_i), with $i = 1, \dots, M$. x_i is the ith input, and y_i is the ith target variable. In this case, the two metrics considered here are the mean squared error (MSE)

$$\text{MSE} = \frac{1}{M}\sum_{i=1}^{M}(y_i - \hat{y}_i)^2 \tag{4}$$

and the mean absolute error (MAE)

$$\text{MAE} = \frac{1}{M}\sum_{i=1}^{M}|y_i - \hat{y}_i| \tag{5}$$

where $\hat{y}_i$ indicates the prediction of the ML model.

A generic **binary classification problem** has the objective of classifying a set of M observations x_i with $i = 1, \dots, M$ into two classes 1 and 0. The class labels to be predicted are indicated, as in the regression problem, with $y_i \in \{0, 1\}$. In this case, the most used metric is the accuracy (a) obtained simply by

$$a = \frac{\text{Number of correctly classified observations}}{M} \tag{6}$$

The impact of unbalanced datasets on the accuracy is not discussed here and the reader is referred to other works (Michelucci, 2018).

The estimation of the expected value and variance of the metrics of Eqs. (4)–(6) that account for the uncertainty ϵ_i on y_i will be given in Sections 3 and 4 for the regression and classification problems respectively.

3. Regression problem

3.1. Mean Squared Error (MSE) estimate

As discussed earlier, the target variables measurements y_i are assumed independent and following a normal distribution (this is a common assumption when dealing with measurement errors) with average $\overline{y}_i$ and standard deviation σ_i (see Eq. (3)). The standard deviation σ_i may differ for different i, for example if y_i has different errors in different ranges of its value. The standard deviation σ_i is a way of giving an estimate of the measurement error.

Let us consider first the special case when $\sigma_i = \sigma$ for $i = 1, \dots, M$ (in other words, when the standard deviation of the y_i is constant). In this case, there is an elegant way to understand everything about the MSE behaviour without any complex calculations. This is reported in Section 3.1.1. The general formulas that apply for σ_i not constant are given in Section 3.1.2.

3.1.1. Constant variances σ_i^2

This section covers the case where $\sigma_i = \sigma$ for $i = 1, \dots, M$. Let us rewrite the MSE as

$$\text{MSE} = \frac{\sigma^2}{M}\sum_{i=1}^{M}\left(\frac{\delta_i}{\sigma}\right)^2 \tag{7}$$

where $\delta_i \equiv y_i - \hat{y}_i$. Note that since δ_i are normally distributed, clearly

$$\frac{\delta_i}{\sigma} \sim \mathcal{N}\left(\frac{\overline{\delta}_i}{\sigma}, 1\right) \tag{8}$$

Eq. (7) is the sum of normally distributed random variables squared with a variance of one, and, therefore, their sum follows the non-central chi-squared distribution with M degrees of freedom (indicated here with $\chi'_M(\lambda)$) (Hogg, Tanis, & Zimmerman, 2010)

$$\sum_{i=1}^{M}\left(\frac{\delta_i}{\sigma}\right)^2 \sim \chi'_M(\lambda) \tag{9}$$

with the noncentrality parameter λ given by

$$\lambda = \sum_{i=1}^{M}\frac{\overline{\delta}_i^2}{\sigma^2} \tag{10}$$

where $\overline{\delta}_i = \overline{y}_i - \hat{y}_i$ Thus from Eqs. (7) and (9) we can say that the MSE satisfies

$$\frac{M}{\sigma^2}\text{MSE} \sim \chi'_M(\lambda). \tag{11}$$

After knowing this, it is straightforward to evaluate

$$\mathbb{E}\left(\frac{M}{\sigma^2}\text{MSE}\right) \tag{12}$$

and

$$\text{Var}\left(\frac{M}{\sigma^2}\text{MSE}\right) \tag{13}$$

In fact, it is a well-known result that for a random variable X that follows a noncentral chi-square distribution $\chi'_M(\lambda)$ (Hogg et al., 2010) it is true that

$$\mathbb{E}(X) = M + \lambda \tag{14}$$

and

$$\text{Var}(X) = 2M + 4\lambda. \tag{15}$$

With the help of Eqs. (11) and (15) the expected value of the MSE can be rewritten in a compact and quite interpretable form:

$$\mathbb{E}\left(\frac{M}{\sigma^2}\text{MSE}\right) = M + \lambda \tag{16}$$

and, therefore,

$$\mathbb{E}(\text{MSE}) = \frac{1}{M}\sum_{i=1}^{M}(\overline{y}_i - \hat{y}_i)^2 + \sigma^2. \tag{17}$$

Eq. (17) indicates that a better estimation of the expected value is obtained by the MSE evaluated with the average of the labels plus variance of the measurements y_i.

The formula for the variance of the MSE is given, using Eq. (15), by the formula

$$\text{Var}(\text{MSE}) = \frac{2\sigma^4}{M} + \frac{4\sigma^2}{M^2}\sum_{i=1}^{M}\overline{\delta}_i^2. \tag{18}$$

Note that the formulas given are only valid in the case where the variances of the single y_i are equal to a constant σ. This is not always the case, and especially in real life cases, quantities may have different measurement errors depending on their values. The general case is discussed in the next section.

3.1.2. Non-constant variances σ_i^2

To evaluate the MSE expected value and its variance in the case where the σ_i are all different, one can use two approaches: a statistical one and an a-priori one that consist in evaluating the necessary integrals directly. The statistical approach is described in this section. The a priori in Appendix A. For non-constant variances, it is impossible to reduce the sum

$$\frac{1}{M}\sum_{i=1}^{M}\delta_i^2 \tag{19}$$

to a sum of M variables with different averages but unit variances (the prerequisites to get the noncentral chi-square distribution used in the previous section). In fact, in this case, in general

$$\delta_i \sim \mathcal{N}\left(\overline{\delta}_i, \sigma_i^2\right) \tag{20}$$

and, therefore, we cannot use the same strategy that was used in the previous section. To determine the expected value and variance, let us observe that

$$\frac{\delta_i}{\sigma_i} \sim \mathcal{N}\left(\frac{\overline{\delta}_i}{\sigma_i}, 1\right) \tag{21}$$

thus

$$\left(\frac{\delta_i}{\sigma_i}\right)^2 \sim \chi'_1(\lambda) \tag{22}$$

with $\lambda = \overline{\delta}_i^2/\sigma_i^2$ is the noncentrality parameter. Using the expected value formula for a non-central chi-squared distribution with one degree of freedom the expectation value is

$$\mathbb{E}\left[\left(\frac{\delta_i}{\sigma_i}\right)^2\right] = 1 + \frac{\overline{\delta}_i^2}{\sigma_i^2} \quad \Rightarrow \quad \mathbb{E}(\delta_i^2) = \sigma_i^2 + \overline{\delta}_i^2 \tag{23}$$

Now

$$\mathbb{E}(\text{MSE}) = \mathbb{E}\left(\frac{1}{M}\sum_{i=1}^{M}\delta_i^2\right) \tag{24}$$

and substituting Eq. (23) in Eq. (24)

$$\begin{aligned}\mathbb{E}(\text{MSE}) = \mathbb{E}\left(\frac{1}{M}\sum_{i=1}^{M}\delta_i^2\right) = \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}(\delta_i^2) = \\ = \frac{1}{M}\sum_{i=1}^{M}(\overline{\delta}_i^2 + \sigma_i^2)\end{aligned} \tag{25}$$

that is the generalized version of Eq. (17). Let us turn our attention to the variance. From Eq. (15) with $M = 1$ it follows that

$$\text{Var}\left(\frac{\delta_i^2}{\sigma_i^2}\right) = 2 + 4\frac{\overline{\delta}_i^2}{\sigma_i^2} \tag{26}$$

and by using the property that

$$\text{Var}\left(\frac{\delta_i^2}{\sigma_i^2}\right) = \frac{1}{\sigma_i^4}\text{Var}(\delta_i^2) \tag{27}$$

the variance becomes

$$\text{Var}(\delta_i^2) = 2\sigma_i^4 + 4\overline{\delta}_i^2\sigma_i^2 \tag{28}$$

and thus since

$$\text{Var}(\text{MSE}) = \text{Var}\left(\frac{1}{M}\sum_{i=1}^{M}\delta_i^2\right) = \frac{1}{M^2}\sum_{i=1}^{M}\text{Var}(\delta_i^2) \tag{29}$$

using Eq. (28)

$$\text{Var}(\text{MSE}) = \frac{2}{M^2}\sum_{i=1}^{M}\sigma_i^4 + \frac{4}{M^2}\sum_{i=1}^{M}\overline{\delta}_i^2\sigma_i^2 \tag{30}$$

that is the generalized version of Eq. (18).

The a priori determination of $\mathbb{E}(\text{MSE})$ is performed by evaluating the integral

$$\mathbb{E}(\text{MSE}) = \frac{1}{M}\sum_{i=1}^{M}\left[\frac{1}{\sqrt{2\pi}\sigma_i}\int_{\mathbb{R}}(y_i - \hat{y}_i)^2 e^{-\frac{(y_i-\overline{y}_i)^2}{2\sigma_i^2}}\,dy_i\right] \tag{31}$$

This calculation requires some work and is shown in Appendix A for completeness.

3.2. Mean Absolute Error (MAE) estimate

Let us turn our attention to the MAE. Since $y_i \sim \mathcal{N}(\overline{y}_i, \sigma_i^2)$ the quantity

$$|\delta_i| = |y_i - \hat{y}_i| \tag{32}$$

follows a folded normal distribution, indicated here with $\mathcal{F}(\bar{\delta}_i, \sigma_i^2)$. The expected value of a random variable X following a folded distribution, $X \sim \mathcal{F}(\mu, \sigma^2)$ is given by Hogg et al. (2010)

$$\mathbb{E}(X) = \mu \sqrt{\frac{2}{\pi}} e^{-\mu^2/(2\sigma^2)} + \mu \left(1 - 2\Phi\left(-\frac{\mu}{\sigma}\right)\right) \tag{33}$$

where Φ is the normal cumulative distribution function. The same formula can be expressed in terms of the error function as

$$\mathbb{E}(X) = \mu \sqrt{\frac{2}{\pi}} e^{-\mu^2/(2\sigma^2)} + \mu \operatorname{erf}\left(\frac{\mu}{\sqrt{2}\sigma}\right). \tag{34}$$

The expected value of MAE is given by

$$\mathbb{E}(\text{MAE}) = \mathbb{E}\left(\frac{1}{M}\sum_{i=1}^{M} |\delta_i|\right) = \frac{1}{M}\sum_{i=1}^{M} \mathbb{E}(|\delta_i|) \tag{35}$$

the expected value is given by

$$\begin{aligned} \mathbb{E}(\text{MAE}) \quad = \quad & \frac{1}{M}\sum_{i=1}^{M} |\hat{y}_i - \bar{y}_i| + \frac{1}{M}\sum_{i=1}^{M} \left\{ \frac{\sqrt{2}}{\sqrt{\pi}} \sigma_i e^{-\delta_i^2/(2\sigma_i^2)} + \right. \\ & \left. - \ |\delta_i| \operatorname{erfc}\left(\frac{|\delta_i|}{\sqrt{2}\delta_i}\right)\right\} \end{aligned} \tag{36}$$

where Eq. (34) has been rewritten using the function erfc() to write $\mathbb{E}(\text{MAE})$ as the sum of the known formula for MAE (albeit evaluated with the averages of y_i) plus a correction term $\Delta(MAE)$

$$\mathbb{E}(\text{MAE}) = \frac{1}{M}\sum_{i=1}^{M} |\hat{y}_i - \bar{y}_i| + \Delta(MAE) \tag{37}$$

with

$$\Delta(\text{MAE}) = \frac{1}{M}\sum_{i=1}^{M} \left\{ \frac{\sqrt{2}}{\sqrt{\pi}} \sigma_i e^{-\delta_i^2/(2\sigma_i^2)} - |\delta_i| \operatorname{erfc}\left(\frac{|\delta_i|}{\sqrt{2}\delta_i}\right)\right\} \tag{38}$$

Finally, the variance of the MAE Var(MAE) is analysed. The variance of a random variable X such that $X \sim \mathcal{F}(\mu, \sigma^2)$ is given by

$$\text{Var}(X) = \mu^2 + \sigma^2 - \overline{X}^2 \tag{39}$$

Considering $X = |\delta_i|$

$$\text{Var}(|\delta_i|) = \bar{\delta}_i^2 + \sigma_i^2 - \left(\bar{\delta}_i \sqrt{\frac{2}{\pi}} e^{-\bar{\delta}_i^2/(2\sigma^2)} - \bar{\delta}_i \operatorname{erf}\left(\frac{\bar{\delta}_i}{\sqrt{2}\sigma}\right)\right)^2 \tag{40}$$

Since the different measurements i are independent, the property $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ can be used (in fact, in this case $\text{Cov}(X, Y) = 0$). Thus, by using Eq. (40) one can derive the following formula

$$\text{Var}(\text{MAE}) = \frac{1}{M^2}\sum_{i=0}^{M} \left\{ \bar{\delta}_i^2 + \sigma_i^2 - \left(\bar{\delta}_i \sqrt{\frac{2}{\pi}} e^{-\bar{\delta}_i^2/(2\sigma^2)} - \bar{\delta}_i \operatorname{erf}\left(\frac{\bar{\delta}_i}{\sqrt{2}\sigma}\right)\right)^2 \right\} \tag{41}$$

4. Classification problem

In a binary classification problem the machine learning model typically outputs the probability $\hat{y}_i$ of an observation of being in class 1 (the labels are $y_i \in \{0, 1\}$). The conversion of the probability into a class is then done using the Heaviside step function $H(x - \alpha)$ where α is a threshold that is normally chosen as $\alpha = 1/2$. The accuracy a can be written for a binary problem in the following form

$$a = \frac{1}{M}\sum_{i=1}^{M} \left[y_i H(\hat{y}_i - \alpha) + (1 - y_i)(1 - H(\hat{y}_i - \alpha))\right] \tag{42}$$

The two terms appearing in Eq. (42) correspond to the true positives (TP) and true negatives (TN). In facts

$$\begin{aligned} \text{TP} \quad &= \quad \frac{1}{M}\sum_{i=1}^{M} \left[y_i H(\hat{y}_i - \alpha)\right] \\ \text{TN} \quad &= \quad \frac{1}{M}\sum_{i=1}^{M} \left[(1 - y_i)(1 - H(\hat{y}_i - \alpha))\right]. \end{aligned} \tag{43}$$

Let us consider the case where there is a probability $q < 1$ that an observation label y_i is wrong. Let us start by considering b_j, a Bernoulli random variable with a probability p of being one (and consequently a probability $q = 1 - p$ of being 0). Let us define the random variable

$${}_r y_i = y_i b_j + (1 - b_j)(1 - y_i). \tag{44}$$

${}_r y_i$ will assume the value y_i with a probability p. ${}_r y_i$ will assume the value of $1 - y_i$ with a probability q. The expectation value and the variance of ${}_r y_i$ are given by

$$\begin{aligned} \mathbb{E}({}_r y_i) =& y_i \mathbb{E}(b_j) + (1 - y_i)\mathbb{E}(1 - b_j) = y_i p + (1 - y_i)q = \\ =& y_i(1 - 2q) + q \end{aligned} \tag{45}$$

and

$$\begin{aligned} \text{Var}({}_r y_i) =& y_i^2 \text{Var}(b_j) + (1 - y_i)^2 \text{Var}(1 - b_j) = \\ =& y_1 pq + (1 - y_i)pq = pq \end{aligned} \tag{46}$$

Eq. (46) can be derived by noting that since $y_i \in \{0, 1\}$ it is true that $y_i^2 = y_i$ and $(1 - y_i)^2 = (1 - y_i)$. The accuracy in the presence of errors in the labels is obtained by using the random variable ${}_r y_i$ in Eq. (42), which results in

$${}_r a = \frac{1}{M}\sum_{i=1}^{M} \left[{}_r y_i H(\hat{y}_i - \alpha) + (1 -_r y_i)(1 - H(\hat{y}_i - \alpha))\right] \tag{47}$$

Now all ingredients are available to calculate $\mathbb{E}({}_r a)$ and $\text{Var}({}_r a)$ with the help of Eqs. (45) and (46). Using the properties of the expected value and of the variance, the following results can be obtained in just a few steps

$$\begin{cases} \mathbb{E}({}_r a) = a + q(1 - 2a) \\ \text{Var}({}_r a) = pq = (1 - q)q. \end{cases} \tag{48}$$

In Appendix C an alternative and more intuitive way of obtaining $\mathbb{E}({}_r a)$ is described.

Eq. (48) is a very interesting result that needs some discussion. First of all, it can be observed that for any model for which $a > 1/2$, $\mathbb{E}({}_r a) < a$, as expected. The errors of the labels effectively reduce the performance of the model. Note that any model in a binary classification problem that has $a < 1/2$ can be transformed into one with $a > 1/2$ by simply exchanging all predictions: 1 into 0 and vice versa.

The expected value of the accuracy, can also be written in a more compact form as

$$\mathbb{E}({}_r a) = \frac{1}{M}\sum_{i=1}^{M} \left[(1 - q)\mathcal{A}(y_i) + q\mathcal{A}(1 - y_i)\right] \tag{49}$$

where $\mathcal{A}(y_i)$ is

$$\mathcal{A}(y_i) = y_i H(\hat{y}_i - \alpha) + (1 - y_i)(1 - H(\hat{y}_i - \alpha)) \tag{50}$$

from Eq. (47). To better understand this formula, one needs to rewrite it a slightly different form by using Eq. (C.5) derived in Appendix B, reported here for clarity

$$\mathbb{E}({}_r a) = (1 - q)\frac{\text{TP} + \text{TN}}{M} + q\frac{\text{FP} + \text{FN}}{M} \tag{51}$$

This formula can be interpreted with the help of Fig. 1:

- If an observation is a true positive or a true negative and the label is wrong, it will be classified with a probability q as either a false positive or false negative, respectively. In other words, the error of the label will reduce the number on the diagonal of the confusion matrix and will contribute to the off-diagonal terms.

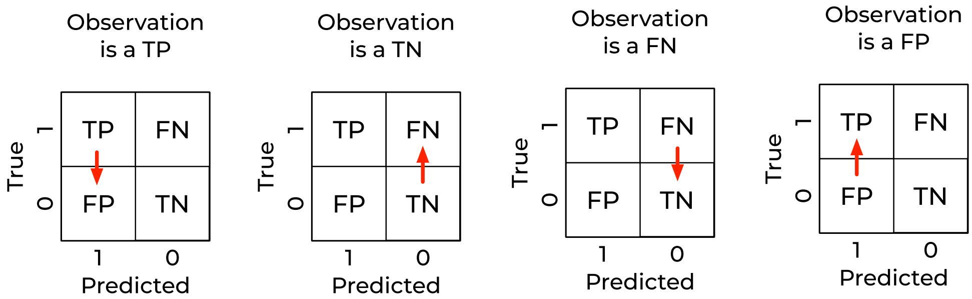

Fig. 1. A visual representation of the effect of a wrong label for an observation: the red arrow indicates how the observation will be classified in the confusion matrix. TP: true positives, TN: true negatives, FP: false positives and FN: false negatives.

- Analogously, if an observation is a false negative or false positive and the label is wrong, the error on the label has the effect of reducing the off-diagonal term of the confusion matrix and contributes to the diagonal terms.

Eq. (48) gives an estimate of accuracy taking into account a possible error in the labels, due for example to measurement errors. The simple accuracy a obtained neglecting measurement errors is an overestimation and does not give a correct picture of how a model could perform.

Let us consider for example a hypothetical model that has obtained $a = 0.85$ with labels with a probability of 5% of being wrong. Eq. (48) gives an estimate of the accuracy of $\mathbb{E}(_r a) = 0.815\%$. The difference is not negligible and must be taken into account for any application of machine learning to scientific results that involves measurement errors (basically always).

5. How to use the formulas

Tables 1 and 2 summarize the formulas derived in this paper. Note that the application of formulas requires not only the standard deviation of the measurements σ_i, but also its average $\overline{y}_i$. When measuring a quantity y_i the scientist should try to get enough measurements to be able to estimate average and standard deviation. Measurements may be expensive, and therefore could lead to having only a limited number of them, making estimating average and standard deviation difficult or, in extreme cases, impossible with a statistical approach. In such a case, the standard deviation should be replaced with the measurement error obtained by a classical propagation of the experimental errors. The average should be evaluated by calculating the mean of the few available values of y_i, or, in the extreme case where only one value is available, using this value in place of the average.

When calculating the MSE, MAE or accuracy the following process should be followed:

1. Multiple measurements for the y_i should be performed to be able to evaluate $\overline{y_i}$. σ_i can be estimated statistically (by evaluating the variance of the multiple measurements of y_i) or by doing error propagation by using the knowledge about how the measurement were performed. The latter way is more practical. In fact to get a good estimate of errors by statistical means one needs a large amount of measurements, and that is not always possible.
2. The appropriate metric from Table 1 (depending on the kind of problem one is trying to solve) should be chosen and calculated according to the formula.
3. If needed, the variance corresponding to the chosen metric should be chosen from Table 2.

The use of the formulas described in this paper will give a more realistic estimate of machine learning metrics (here MSE, MAE and accuracy) and therefore of the model performance since it takes into account measurement errors on the target variables.

Table 1
Formulas for the expected value of the MSE, MAE and accuracy (a) that take into account errors on the target variables.

Metric	Formula
MSE	$\mathbb{E}(\text{MSE}) = \frac{1}{M}\sum_{i=1}^{M}(\overline{y}_i - \hat{y}_i)^2 + \sigma^2$ Case when $\sigma_i = \sigma$ for $i = 1, \dots, M$
MSE	$\mathbb{E}(\text{MSE}) = \frac{1}{M}\sum_{i=1}^{M}((\overline{y}_i - \hat{y}_i)^2 + \sigma_i^2)$ For σ_i not constant
MAE	$\mathbb{E}(\text{MAE}) = \frac{1}{M}\sum_{i=1}^{M}\lvert\hat{y}_i - \overline{y}_i\rvert +$ $\frac{1}{M}\sum_{i=1}^{M}\left\{\frac{\sqrt{2}}{\sqrt{\pi}}\sigma_i e^{-\delta_i^2/(2\sigma_i^2)} - \lvert\delta_i\rvert \text{erfc}\left(\frac{\lvert\delta_i\rvert}{\sqrt{2}\delta_i}\right)\right\}$ For σ_i not constant
Accuracy	$\mathbb{E}(_r a) = a + q(1 - 2a)$

Table 2
Formulas for the variance of the MSE, MAE and accuracy (a).

Metric	Formula
MSE	$\text{Var}\,(\text{MSE}) = \frac{2}{M^2}\sum_{i=1}^{M}\sigma_i^4 + \frac{4}{M^2}\sum_{i=1}^{M}\overline{\delta}_i^2\sigma_i^2$
MAE	$\text{Var}(\text{MAE}) = \frac{1}{M^2}\sum_{i=0}^{M}\left\{\overline{\delta}_i^2 + \sigma_i^2 - \left(\overline{\delta}_i\sqrt{\frac{2}{\pi}}e^{-\overline{\delta}_i^2/(2\sigma^2)} - \overline{\delta}_i \text{erf}\left(\frac{\overline{\delta}_i}{\sqrt{2}\sigma}\right)\right)^2\right\}$
Accuracy	$\text{Var}(_r a) = pq = q(1 - q)$

It is important to discuss the applicability of the formulas derived in this paper. Firstly, the formulae are based on the assumption that the errors on labels are distributed according to a Gaussian distribution. This is not always the case, as can be seen in various cases as for example in Astrophysics (Chen, Gott III, & Ratra, 2003) or degradation analysis (Zhai & Ye, 2017). The interested reader is referred to the review by Bailey (2017). In their work, the authors analyse different cases in medicine and physics and highlight how error distributions deviate in some cases from a Gaussian distribution and are more close to a Cauchy distribution.

Secondly, the formulas are applicable when the measurements are independent (see Section 2). This is of course not always applicable, as spatio-temporal correlations may be present. In such a case, the covariance matrix is non-diagonal. Nevertheless, the assumption of the independence of the measurements is often a good approximation in many practical cases, thus making the derived formula a good approximation of the expected value and the variance.

Lastly, it should be noted that, although there is no assumptions on the value of M, the formulae are most useful when applied to classical machine learning use-case where $M \gg 1$. If a set of measurements consists of very few values, it is questionable how machine learning could be applied.

6. Conclusions

This work presents for the first time formulas for calculating the metrics commonly used in ML, namely MSE, MAE, and accuracy, taking into account the errors in the target variables. The formulas, which are of general validity, are derived using both a statistical and an a priori approach. They give more realistic estimates of the metrics that are otherwise overly optimistic. The analysis shows that the MSE and MAE calculated with the derived formulas are always larger than the one obtained ignoring errors in the measurements (in other words, setting $\sigma_i = 0$ for $i = 1, \dots, M$). The accuracy evaluated according to the formula given in Table 1 is always lower than the one evaluated by using only the target variables and ignoring possible errors.

Another important contribution of this paper is that it shows the relevance of performing multiple repeated measurements to calculate averages and variances of measurements. These are crucial to obtain scientifically accurate estimates of ML metrics, and therefore, ML model performances. The reported formulas have a very wide applicability and should be used any time the target variables are known within an error.

CRediT authorship contribution statement

Umberto Michelucci: Supervision, Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Formal analysis. **Francesca Venturini:** Methodology, Writing – review & editing, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Appendix A. Direct calculation of $\mathbb{E}$(MSE)

The integral to be evaluated is

$$\mathbb{E}(\mathrm{MSE}) = \frac{1}{M}\sum_{i=1}^{M}\left[\frac{1}{\sqrt{2\pi}\sigma_i}\int_{\mathbb{R}}(y_i-\hat{y}_i)^2 e^{-\frac{(y_i-\overline{y}_i)^2}{2\sigma_i^2}}\,dy_i\right] \tag{A.1}$$

To solve this integral, the following change of variable can be used

$$s = \frac{y_i-\overline{y}_i}{\sqrt{2}\sigma_i} \tag{A.2}$$

this of course leads to

$$dy_i = ds\sqrt{2}\sigma_i. \tag{A.3}$$

Therefore, Eq. (A.1) can be rewritten as

$$\frac{1}{\sqrt{\pi}M}\sum_{i=1}^{M}\int_{\mathbb{R}}(s\sigma_i\sqrt{2}+\overline{y}_i-\hat{y}_i)^2 e^{-s^2}\,ds \tag{A.4}$$

and by expanding the polynomial squared and defining $\delta_i = \hat{y}_i - \overline{y}_i$ one obtains

$$\frac{1}{\sqrt{\pi}M}\sum_{i=1}^{M}\int_{\mathbb{R}}(2s^2\sigma_i^2+\delta_i^2-2\sqrt{2}\sigma_i s\delta_i)e^{-s^2}\,ds \tag{A.5}$$

in Eq. (A.5) there are three terms that need to be evaluated.

$$\mathbb{E}(\mathrm{MSE}) = A + B + C \tag{A.6}$$

with

$$\begin{aligned} A &= \frac{2}{\sqrt{\pi}M}\sum_{i=1}^{M}\sigma_i^2\int_{\mathbb{R}}s^2e^{-s^2}ds \\ B &= \frac{1}{\sqrt{\pi}M}\sum_{i=1}^{M}\delta_i^2\int_{\mathbb{R}}e^{-s^2}ds \\ C &= -\frac{2\sqrt{2}}{\sqrt{\pi}M}\sum_{i=1}^{M}\sigma_i\delta_i\int_{\mathbb{R}}se^{-s^2}ds \end{aligned} \tag{A.7}$$

given the symmetry of the function under the integral sign in C it is immediately evident that $C = 0$. Using the results,

$$\int_{\mathbb{R}}e^{-s^2}ds = \sqrt{\pi} \tag{A.8}$$

and

$$\int_{\mathbb{R}}s^2e^{-s^2}ds = \frac{\sqrt{\pi}}{2} \tag{A.9}$$

A and B can be easily calculated

$$\begin{aligned} A &= \frac{1}{M}\sum_{i=1}^{M}\sigma_i^2 \\ B &= \frac{1}{M}\sum_{i=1}^{M}\delta_i^2 \end{aligned} \tag{A.10}$$

Note how B is the MSE evaluated with the measurement averages $\overline{y}_i$, while A is the average of the measurement standard deviations. So Eq. (A.5) can be finally rewritten as

$$\mathbb{E}(\mathrm{MSE}) = \frac{1}{M}\sum_{i=1}^{M}\sigma_i^2 + \frac{1}{M}\sum_{i=1}^{M}(\overline{y}_i-\hat{y}_i)^2 \tag{A.11}$$

This concludes the derivation. The calculation of Var(MSE) is not reported here, as it is similar to the one for the expected value and would make this paper unbearably long.

Appendix B. Direct calculation of $\mathbb{E}$(MAE)

The integral to be evaluated is

$$\mathbb{E}(\mathrm{MAE}) = \frac{1}{M}\sum_{i=1}^{M}\underbrace{\left[\frac{1}{\sqrt{2\pi}\sigma_i}\int_{\mathbb{R}}|y_i-\hat{y}_i|e^{-\frac{(y_i-\overline{y}_i)^2}{2\sigma_i^2}}\,dy_i\right]}_{J} \tag{B.1}$$

Let us consider for the calculation only J. The following change of variables can be used

$$s = \frac{y_i-\overline{y}_i}{\sqrt{2}\sigma_i} \rightarrow ds = dy_i\frac{1}{\sqrt{2}\sigma_i} \tag{B.2}$$

therefore

$$\begin{aligned} J &= \int_{\mathbb{R}}|s\sqrt{2}\sigma_i - \underbrace{(\hat{y}_i-\overline{y}_i)}_{\delta_i}|e^{-s^2}ds = \\ &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}|s\sqrt{2}\sigma_i-\delta_i|e^{-s^2}ds \end{aligned} \tag{B.3}$$

due to the absolute value, the integral must be split into two parts: $J = J_A + J_B$. Part A for $s\sqrt{2}\sigma_i - \delta_i \geq 0$ and part B for $s\sqrt{2}\sigma_i - \delta_i < 0$. The two integrals are

$$J_A = \frac{\sqrt{2}\sigma_i}{\sqrt{\pi}}\int_{(\delta_i/(\sqrt{2}\sigma_i))}^{\infty}\left(s-\frac{\delta_i}{\sqrt{2}\sigma_i}\right)e^{-s^2}ds \tag{B.4}$$

and

$$J_B = -\frac{\sqrt{2}\sigma_i}{\sqrt{\pi}}\int_{\infty}^{(\delta_i/(\sqrt{2}\sigma_i))}\left(s-\frac{\delta_i}{\sqrt{2}\sigma_i}\right)e^{-s^2}ds \tag{B.5}$$

Let us start with J_A. To further simply the notation let us define

$$\tilde{\delta}_i = \frac{\delta_i}{\sqrt{2}\sigma_i} \tag{B.6}$$

so J_A can now be evaluated

$$\begin{aligned} J_A &= \frac{\sigma_i\sqrt{2}}{\sqrt{\pi}}\left[\int_{\tilde{\delta}_i}^{\infty} se^{-s^2}ds - \int_{\tilde{\delta}_i}^{\infty}\tilde{\delta}_i e^{-s^2}ds\right] \\ &= \frac{\sigma_i\sqrt{2}}{\sqrt{\pi}}\left[\frac{1}{2}e^{-\tilde{\delta}_i^2} - \tilde{\delta}_i\frac{\sqrt{\pi}}{2}\mathrm{erfc}(\tilde{\delta}_i)\right] \\ &= \frac{\sigma_i}{\sqrt{2\pi}}e^{-\tilde{\delta}_i^2} - \frac{1}{\sqrt{2}}\delta_i\mathrm{erfc}(\tilde{\delta}_i) \end{aligned} \tag{B.7}$$

Analogously

$$\begin{aligned} J_B &= -\frac{\sigma_i\sqrt{2}}{\sqrt{\pi}}\left[\int_{\infty}^{\tilde{\delta}_i} se^{-s^2}ds - \int_{\infty}^{\tilde{\delta}_i}\tilde{\delta}_i e^{-s^2}ds\right] \\ &= \frac{\sigma_i\sqrt{2}}{\sqrt{\pi}}\left[\frac{1}{2}e^{-\tilde{\delta}_i^2} + \tilde{\delta}_i\frac{\sqrt{\pi}}{2}\mathrm{erfc}(-\tilde{\delta}_i)\right] \\ &= \frac{\sigma_i}{\sqrt{2\pi}}e^{-\tilde{\delta}_i^2} + \frac{1}{\sqrt{2}}\delta_i\mathrm{erfc}(-\tilde{\delta}_i) \end{aligned} \tag{B.8}$$

therefore, with some simplifications

$$J = \frac{\sqrt{2}\sigma_i}{\sqrt{\pi}}\left[e^{-\tilde{\delta}_i^2} + \sqrt{\pi}\tilde{\delta}_i\mathrm{erf}(\tilde{\delta}_i)\right] \tag{B.9}$$

Now this form is not easy to interpret, and it can be brought in a more interpretable form with some additional manipulation. Let us start by noticing that

$$\tilde{\delta}_i\mathrm{erf}(\tilde{\delta}_i) = |\tilde{\delta}_i|\mathrm{erf}(|\tilde{\delta}_i|) \tag{B.10}$$

since $\tilde{\delta}_i\mathrm{erf}(\tilde{\delta}_i) = -\tilde{\delta}_i\mathrm{erf}(-\tilde{\delta}_i)$ due to the fact that $\mathrm{erf}(-x) = -\mathrm{erf}(x)$. Additionally, Eq. (B.10) can be rewritten as

$$\tilde{\delta}_i\mathrm{erf}(\tilde{\delta}_i) = |\tilde{\delta}_i|\mathrm{erf}(|\tilde{\delta}_i|) = |\tilde{\delta}_i|(1 - \mathrm{erfc}(|\tilde{\delta}_i|)) \tag{B.11}$$

where $\mathrm{erfc}(x)$ is the complementary error function. Using Eq. (B.11), Eq. (B.9) can be rewritten as

$$J = |\delta_i| + \frac{\sqrt{2}}{\sqrt{\pi}}\sigma_i e^{-\tilde{\delta}_i^2} - |\delta_i|\mathrm{erfc}\left(\frac{|\delta_i|}{\sqrt{2}\sigma_i}\right) \tag{B.12}$$

Now $\mathbb{E}(\mathrm{MAE})$ can be finally written

$$\begin{aligned} \mathbb{E}(\mathrm{MAE}) &= \frac{1}{M}\sum_{i=1}^{M}|\hat{y}_i - \overline{y}_i| + \frac{1}{M}\sum_{i=1}^{M}\left\{\frac{\sqrt{2}}{\sqrt{\pi}}\sigma_i e^{-\delta_i^2/(2\sigma_i^2)}\right. \\ &\quad \left. - \ |\delta_i|\mathrm{erfc}\left(\frac{|\delta_i|}{\sqrt{2}\delta_i}\right)\right\} \end{aligned} \tag{B.13}$$

This concludes the derivation.

Appendix C. Alternative calculation of $\mathbb{E}({}_r a)$

The starting point of this alternative derivation is the formula

$$\mathbb{E}({}_r a) = \frac{1}{M}\sum_{i=1}^{M}\left[(1-q)\mathcal{A}({}_r y_i) + q\mathcal{A}(1 -_r y_i)\right] \tag{C.1}$$

where $\mathcal{A}({}_r y_i)$ is

$$\mathcal{A}({}_r y_i) =_r y_i H(\hat{y}_i - \alpha) + (1 -_r y_i)(1 - H(\hat{y}_i - \alpha)) \tag{C.2}$$

from Eq. (47). Eq. (C.1) can be expanded by using Eq. (C.2) as

$$\begin{aligned} \mathbb{E}({}_r a) &= \frac{1}{M}\sum_{i=1}^{M}\left[(1-q)_r y_i H(\hat{y}_i - \alpha) + (1 -_r y_i)(1 - H(\hat{y}_i - \alpha)) + \right. \\ &\quad \left. q(1 -_r y_i)H(\hat{y}_i - \alpha) +_r y_i(1 - H(\hat{y}_i - \alpha))\right] \\ &= \frac{1}{M}\sum_{i=1}^{M}\left[{}_r y_i H(\hat{y}_i - \alpha) - q\ {}_r y_i H(\hat{y}_i - \alpha) + \right. \\ &\quad +(1 -_r y_i)(1 - H(\hat{y}_i - \alpha)) + \\ &\quad -q(1 -_r y_i)(1 - H(\hat{y}_i - \alpha)) + q(1 -_r y_i)H(\hat{y}_i - \alpha) + \\ &\quad \left. +q\ {}_r y_i(1 - H(\hat{y}_i - \alpha))\right] \end{aligned} \tag{C.3}$$

After some algebra, this can be simplified and brought in the form

$$\begin{aligned} \mathbb{E}({}_r a) &= \frac{1}{M}\sum_{i=1}^{M}\left[(1-q)\underbrace{(1 -_r y_i)(1 - H(\hat{y}_i - \alpha))}_{\mathrm{TN}} + \right. \\ &+ \quad (1-q)\underbrace{{}_r y_i H(\hat{y}_i - \alpha)}_{\mathrm{TP}} + \\ &+ \quad \left. q\underbrace{(1 -_r y_i)H(\hat{y}_i - \alpha)}_{\mathrm{FP}} + q\underbrace{{}_r y_i(1 - H(\hat{y}_i - \alpha))}_{\mathrm{FN}}\right] \end{aligned} \tag{C.4}$$

In Eq. (C.4) it is clearly indicated which part gives (when summed) the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). So Eq. (C.4) can be rewritten as

$$\mathbb{E}({}_r a) = (1-q)\frac{\mathrm{TP}+\mathrm{TN}}{M} + q\frac{\mathrm{FP}+\mathrm{FN}}{M} \tag{C.5}$$

The final formula can be easily obtained by noting that

$$\begin{aligned} a &= \frac{\mathrm{TP}+\mathrm{TN}}{M} \\ 1-a &= \frac{\mathrm{FP}+\mathrm{FN}}{M} \end{aligned} \tag{C.6}$$

This concludes the derivation.

References

Bahri, D., Jiang, H., & Gupta, M. (2020). Deep k-NN for noisy labels. In H. D. III, & A. Singh (Eds.), *Proceedings of machine learning research: vol. 119, Proceedings of the 37th International conference on machine learning* (pp. 540–550). PMLR.

Bailey, D. C. (2017). Not Normal: the uncertainties of scientific measurements. *Royal Society Open Science*, *4*(1), Article 160600.

Baldi, P., Bauer, K., Eng, C., Sadowski, P., & Whiteson, D. (2016). Jet substructure classification in high-energy physics with deep neural networks. *Physical Review D*, *93*(9), Article 094034.

Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., et al. (2019). Machine learning and the physical sciences. *Reviews of Modern Physics*, *91*(4), Article 045002.

Carrasco Kind, M., & Brunner, R. J. (2013). TPZ: photometric redshift PDFs and ancillary information by using prediction trees and random forests. *Monthly Notices of the Royal Astronomical Society*, *432*(2), 1483–1501.

Carrasquilla, J., & Melko, R. G. (2017). Machine learning phases of matter. *Nature Physics*, *13*(5), 431–434.

Chen, G., Gott III, J. R., & Ratra, B. (2003). Non-Gaussian Error Distribution of Hubble Constant Measurements. *Publications of the Astronomical Society of the Pacific*, *115*(813), 1269–1279. http://dx.doi.org/10.1086/379219, URL arXiv:astro-ph/0308099.

Collister, A., Lahav, O., Blake, C., Cannon, R., Croom, S., Drinkwater, M., et al. (2007). Megaz-LRG: a photometric redshift catalogue of one million SDSS luminous red galaxies. *Monthly Notices of the Royal Astronomical Society*, *375*(1), 68–76.

Cour, T., Sapp, B., & Taskar, B. (2011). Learning from partial labels. *Journal of Machine Learning Research*, *12*, 1501–1536.

Ghosh, A., & Nachman, B. (2022). A cautionary tale of decorrelating theory uncertainties. *The European Physical Journal C*, *82*(1), 1–11.

Guest, D., Collado, J., Baldi, P., Hsu, S.-C., Urban, G., & Whiteson, D. (2016). Jet flavor classification in high-energy physics with deep neural networks. *Physical Review D*, *94*(11), Article 112002.

Hogg, R. V., Tanis, E. A., & Zimmerman, D. L. (2010). *Probability and statistical inference*. Pearson/Prentice Hall Upper Saddle River, NJ, USA:.

Krishnan, N. A., Mangalathu, S., Smedskjaer, M. M., Tandia, A., Burton, H., & Bauchy, M. (2018). Predicting the dissolution kinetics of silicate glasses using machine learning. *Journal of Non-Crystalline Solids*, *487*, 37–45.

Liu, T., & Tao, D. (2016). Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(3), 447–461. http://dx.doi.org/10.1109/TPAMI.2015.2456899.

Luo, X., Lorentzen, R. J., & Bhakta, T. (2021). Accounting for model errors of rock physics models in 4D seismic history matching problems: A perspective of machine learning. *Journal of Petroleum Science and Engineering*, *196*, Article 107961. http://dx.doi.org/10.1016/j.petrol.2020.107961.

Menon, A., Rooyen, B. V., Ong, C. S., & Williamson, B. (2015). Learning from corrupted binary labels via class-probability estimation. In F. Bach, & D. Blei (Eds.), *Proceedings of machine learning research*: *vol. 37*, *Proceedings of the 32nd international conference on machine learning* (pp. 125–134). Lille, France: PMLR.

Michelucci, U. (2018). *Applied deep learning*. Springer.

Michelucci, U., & Venturini, F. (2019). Multi-task learning for multi-dimensional regression: application to luminescence sensing. *Applied Sciences*, *9*(22), 4748.

Michelucci, U., & Venturini, F. (2021). Estimating neural network's performance with bootstrap: A tutorial. *Machine Learning and Knowledge Extraction*, *3*(2), 357–373.

Morningstar, A., & Melko, R. G. (2018). Deep learning the ising model near criticality. *Journal of Machine Learning Research*.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., & Tewari, A. (2013). Learning with noisy labels. *Advances in Neural Information Processing Systems*, *26*, 1196–1204.

de Oliveira, L., Kagan, M., Mackey, L., Nachman, B., & Schwartzman, A. (2016). Jet-images—deep learning edition. *Journal of High Energy Physics*, *2016*(7), 1–32.

Ravanbakhsh, S., Oliva, J., Fromenteau, S., Price, L., Ho, S., Schneider, J., et al. (2016). Estimating cosmological parameters from the dark matter distribution. In *International conference on machine learning* (pp. 2407–2416). PMLR.

Tanaka, A., & Tomiya, A. (2017). Detection of phase transition via convolutional neural networks. *Journal of the Physical Society of Japan*, *86*(6), Article 063001.

Taylor, J. (1997). Introduction to error analysis, the study of uncertainties in physical measurements.

Wilkinson, C. J., Trivelpiece, C., Hust, R., Welch, R. S., Feller, S. A., & Mauro, J. C. (2022). Hybrid machine learning/physics-based approach for predicting oxide glass-forming ability. *Acta Materialia*, *222*, Article 117432.

Yao, Q., Yang, H., Han, B., Niu, G., & Kwok, J. T.-Y. (2020). Searching to exploit memorization effect in learning with noisy labels. In H. D. III, & A. Singh (Eds.), *Proceedings of machine learning research*: *vol. 119*, *Proceedings of the 37th International conference on machine learning* (pp. 10789–10798). PMLR.

Zhai, Q., & Ye, Z.-S. (2017). Robust degradation analysis with non-Gaussian measurement errors. *IEEE Transactions on Instrumentation and Measurement*, *66*(11), 2803–2812.

Zhang, X.-L., Xiao, H., Luo, X., & He, G. (2022). Ensemble Kalman method for learning turbulence models from indirect observation data. *Journal of Fluid Mechanics*, *949*, A26. http://dx.doi.org/10.1017/jfm.2022.744.

Zheng, S., Wu, P., Goswami, A., Goswami, M., Metaxas, D., & Chen, C. (2020). Error-bounded correction of noisy labels. In H. D. III, & A. Singh (Eds.), *Proceedings of machine learning research*: *vol. 119*, *Proceedings of the 37th International conference on machine learning* (pp. 11447–11457). PMLR.

Zhu, Q., Liu, Z., & Yan, J. (2021). Machine learning for metal additive manufacturing: predicting temperature and melt pool fluid dynamics using physics-informed neural networks. *Computational Mechanics*, *67*(2), 619–635.