Addressing misclassification costs in machine learning through asymmetric loss functions

Bryan M. Barnes[a] and Mark-Alexander Henn[b,c]

[a]Nanoscale Device Characterization Division, National Institute of Standards and Technology, 100 Bureau Drive MS 8423, Gaithersburg, MD 20899, USA
[b]Applied Mathematics Division, National Institute of Standards and Technology, 100 Bureau Drive MS 8910, Gaithersburg, MD 20899, USA
[c]College of Computer, Mathematical & Natural Sciences, University of Maryland, 7998 Regents Dr, College Park, MD 20742, USA

ABSTRACT

Background: Patterning defect metrology requires data interpretation with classification, each well-suited to machine learning (ML). Defect classification however has notable misclassification costs; mislabeling a defect as nominal has greater impact than the converse.
Aim: Though quantified costs are not publicly available, total economic misclassification cost (total cost) is optimized across orders-of-magnitude variation in cost ratio $\mathcal{C}$ and classification threshold $0.01 < \tau < 0.99$.
Approach: Convolutional neural networks are trained using the intrinsically weighted and scaled asymmetric focal losses (AFL, sAFL) with hyperparameter γ with weighted and unweighted binary cross-entropy (wBCE, BCE) functions trained for comparisons. Optimal functions and conditions are identified for reducing total cost. For reproducibility, publicly available ML data sets are surrogates for industrial imaging data.
Results: For these data the sAFL mimimizes total cost at $\tau = 0.5$, $\mathcal{C} \geq 16$. The AFL reduces total cost at $0.1 \leq \tau < 0.5$, $\mathcal{C} > 128$. Asymmetric loss functions lower total cost versus wBCE by 15 % to 40 % for $0.2 < \tau < 0.5$, $\mathcal{C} > 64$.
Conclusions: Total economic misclassification cost can be tailored using asymmetric focal losses. Estimations are presented to allow the extension of reported trends to industrial applications with strong class imbalances between defect-indicative and nominal-indicative data.

Keywords: asymmetric loss functions, asymmetric focal loss, goal-oriented metrics, defect metrology, binary classification, machine learning, convolutional neural networks, scaled asymmetric focal loss

1. INTRODUCTION

Patterning errors ("defects") in semiconductor manufacturing affect the production yield as such imperfections may render computer chips electrically inoperable. Defect inspection is a crucial step in microelectronics fabrication performed using various measurement techniques including optical imaging,[1] optical scattering,[2] scanning electron microscopy,[3] and voltage contrast imaging.[4] Defect inspection is notably different from many other types of measurements in semiconductor manufacturing, separating measurement data into classifications for process control as opposed to translating measurements into physical quantities such as line width or line height.

Machine learning (ML) augments automated process control and defect classification. By industrial necessity the defect metrology literature (Sec. 2.2) may withhold information needed for understanding the ML utilized, including access to data sets and information on misclassification costs and class imbalance. Qualitatively there should be orders-of-magnitude more data indicative of nominal patterning in lithography than data indicative of defective patterning, yielding class imbalances. There should also be misclassification cost imbalances; misclassifying a nominal pattern as defective can lead to the inconvenience of scrapping or re-patterning ("reworking") of

Further author information: (Send correspondence to B.M.B.)
B.M.B.: E-mail: bmbarnes@nist.gov, Telephone: 1 301 975 3947
M-A.H..: E-mail: mark.henn@nist.gov, Telephone: 1 301 975 5067

the wafer, but defective patterning that is misclassified may lead to continued, costly fabrication of a device that from its beginnings has been inoperable. Cost-sensitive loss functions are well-studied (Sec. 2.3.1) with some reports also considering the economic ramifications of misclassification and the costs created (Sec. 2.3.2). Without quantitative values for both cost and economic imbalances, developing cost-sensitive strategies for defect detection is inherently challenging. Semiconductor manufacturing metrologists likely possess experimental insights on this class imbalance and may also understand the orders-of-magnitude differences among misclassification costs.

This work investigates the potential of tailoring asymmetric loss functions to reduce total economic misclassification cost (total cost) due to economic imbalance in defect metrology. As one trains a neural network (NN) to improve the network's predicted probabilities $\hat{y}$ relative to the true training values y, it is the loss function that is minimized; the loss function provides the crucial feedback to properly train the NN. Prior knowledge that misclassifications of one class are qualitatively more valuable motivates this application of asymmetric loss functions. One key contribution of this work is a mechanism for minimizing the total economic misclassification cost $\mathcal{R}$ as functions of misclassification cost ratio $\mathcal{C}$ and classification threshold τ in the absence of quantitative industrial misclassification costs. Another key contribution is the application of a scaling term to the asymmetric focal loss (AFL) loss function,[5–8] yielding a scaled AFL (sAFL) with average magnitudes comparable to the binary cross-entropy (BCE). This scaling removes intrinsic average weighting effects of the AFL permitting a clearer understanding of how differences in functional curvature within the loss function can contribute to optimizing total cost.

2. RELATED WORK

2.1 Semiconductor defect metrology

Semiconductor fabrication metrologists are concerned with defects at all stages of production,[9] from the cleanliness of the unpatterned silicon wafer through each of the approximately 500 process steps[10] required to make modern devices. However, it is at the initial patterning layers at which the sizes of "killer" defects (i.e., creating electrically measurable faults)[11] are smallest. Optical methods are extensively used for defect metrology. Challenges to the continued applicability of optical methods for defects less than 10 nm wide have been identified[12] and potential solutions have been reviewed by Zhu *et al.* recently.[13] Multi-beam or highly parallel scanning electron microscopy (SEM) would overcome its inherent speed and field-of-view limitations while providing much higher resolution data.[14] A specific electron scanning mode called voltage contrast[4] can identify electrical shorts earlier in the production flow, which should directly reduce but not eliminate the misclassification cost imbalance addressed in this work.

2.2 Machine learning in semiconductor metrology

Semiconductor fabrication facilities are data-rich environments given the rigorous process control required at each process step, with several uses of ML identified in the literature. Bischoff *et al.* reported using artificial neutral networks (ANNs) in the analysis of optics-based measurements of 0.25 µm line widths, also called critical dimensions (CDs);[15] state-of-the-art line widths are over 10 times smaller. NNs have been implemented for processing optical measurements of photolithographic exposure rates during patterning.[16] Overlay metrology, the measurement of the displacement between subsequent photolithographic layers, is also now aided using NNs.[17]

Section 1 noted that defect metrology yields classifications rather than quantitative values. Furthermore, unlike other classifications based upon external quantitative scalar inputs (e.g., checking and savings account balances as inputs to a CNN predicting mortgage defaults[18]), defect metrology can use ML both to interpret physical measurement data and to enable classification. The Authors examined this for simulated optical imaging of intentional defect array wafers.[19] Others have focused on SEM images of defects using ML methods such as CNNs[20–23] and autoencoders.[24]

2.3 Cost-sensitive loss functions

2.3.1 Class imbalance and the asymmetric focal loss

Cost-sensitive loss functions address one or both of two key challenges: class imbalance and economic imbalance. Most papers with cost-sensitive functions address class imbalance, see review by Johnson *et al.*,[25] Multiple approaches exist to address class imbalance, but this work draws attention to a relatively new approach to this problem, the asymmetric focal loss (AFL). Focal loss (FL) was introduced by Lin *et al.*[26] in 2017 expressed here for two classes,

$$\mathcal{L}_{\mathrm{FL}} = -y(1-\hat{y})^{\gamma}\log(\hat{y}) - (1-y)(\hat{y})^{\gamma}\log(1-\hat{y}), \tag{1}$$

where $y, \hat{y}$ are the true and predicted labels $y \in [0,1]$ and γ is a hyperparameter $\gamma \geq 0$. If $\gamma = 0$, this reduces to the widely used binary cross-entropy (BCE),

$$\mathcal{L}_{\mathrm{BCE}} = -y\log(\hat{y}) - (1-y)\log(1-\hat{y}). \tag{2}$$

The goal of the FL is to underweight easy-to-classify inputs (e.g., input data x labeled at $y = 0$ and reported as $\hat{y} = 0.05$) while increasing specificity near the decision boundary $\tau = 0.5$.

Recent reports have merged the best attributes of the BCE and FL into the AFL, formulated in this work as

$$\mathcal{L}_{\mathrm{AFL}} = -y\log(\hat{y}) - (1-y)(\hat{y})^{\gamma}\log(1-\hat{y}), \tag{3}$$

with the hyperparameter γ applied only to Class 0.

Imoto *et al.* employed the AFL to address a class imbalance between labeled sounds and silence in acoustic data,.[6] Li *et al.*[5] and Vogt *et al.*[7] both utilized the AFL and CNN for image segmentation for medical imaging. Chen *et al.* proposed using two separate hyperparameters $\gamma_{pos}, \gamma_{neg}$ within the focal loss for fault detection, intrinsically presenting greater tunability to the loss function.[8] Note, one potential reason that the AFL is adept at handling class imbalanced data is that not only does the focal loss component resolve difficult misclassifications but also the FL term (for $y = 0$ here) is inherently under-weighted versus the BCE component (for $y = 1$).

2.3.2 Economic class imbalance

Quantification of the consequences among types of misclassification is beyond the realm of ML and is often left to subject matter experts.[27] Researchers addressing misclassification costs can denote costs for each possible misclassification, or for binary classification offer a ratio of the two misclassification costs. For binary classification, the economic costs can be expressed as a "cost matrix" as shown in Table 1. Numerical evaluations require defined values for these symbolic costs. Such values not only allow weights assigned to the loss function for training but also permitting misclassification results to be evaluated for their "real-world" impact[27] or its associated total misclassification cost.[28] Below, the total misclassification cost is defined with a numerical example to better illustrate the interplay among the individual and total misclassification coast and the classification threshold.

Table 1. Cost matrix for binary classification with actual values y and predicted values $\hat{y}$.

	$\hat{y} = 0$	$\hat{y} = 1$
$y = 0$	$\mathcal{C}_{00}$	$\mathcal{C}_{01}$
$y = 1$	$\mathcal{C}_{10}$	$\mathcal{C}_{11}$

3. APPROACH TO MISCLASSIFICATION COSTS

The approach in this work to minimize the total economic misclassification cost without prior knowledge of the individual misclassification costs is introduced using Fig 1 with histograms of distributions of labels from optimized convolutional neural networks (CNNs). Two separate CNNs, one per row, yield predicted values $\hat{y}$ for each input x with known label y. Members of Class 0 with $\hat{y} > \tau$ and members of Class 1 with $\hat{y} < \tau$ are misclassifications. For the top row, Fig. 1(a), the loss function is symmetric while the loss function is asymmetric for the bottom row, Fig. 1(b). The difference between the first and second columns is a shift in classification

Table 2. Confusion matrix for binary classification denoting the number of true negatives TN, false positives FP, true positives TP, and false negatives FN.

	$\hat{y}=0$	$\hat{y}=1$
$y=0$	TN	FP
$y=1$	FN	TP

thresholds, from $\tau = 0.50$ to $\tau = 0.29$. From these four histograms, one can count the classification results in a 2×2 *confusion matrix*, that reports the number of misclassifications as off-diagonal elements. Table 2 shows a more general formulation for reporting the confusion matrix for binary classification where $y = 0$ is the 'negative' case and $y = 1$ is the 'positive' case in this work. Tables 1 and 2 can be combined to determine the total economic misclassification cost,

$$\mathcal{R} = \mathcal{C}_{01} \cdot \mathsf{FP} + \mathcal{C}_{10} \cdot \mathsf{FN}, \tag{4}$$

where the misclassification costs can be restated as $\mathcal{C}_{01} \equiv \mathcal{C}_{\mathsf{FP}}$ and $\mathcal{C}_{10} \equiv \mathcal{C}_{\mathsf{FN}}$ with true classification costs $\mathcal{C}_{00} = \mathcal{C}_{11} \equiv 0$.

Many classification metrics are available including the accuracy,

$$A = \frac{\mathsf{TP} + \mathsf{TN}}{\mathsf{FP} + \mathsf{TN} + \mathsf{TP} + \mathsf{FN}}, \tag{5}$$

using definitions from Table 2. In a recent comparison of 136 possible methods (both ML and non-ML) to solve binary classification problems for economics, Gerunov noted "Problems of binary choice are often connected to high-stakes decisions with potentially large impact, which is why achieving high accuracy is of significant importance."[29] From Eq. 5 and Fig. 1(c), the combination with the highest accuracy is for the upper left-hand panel using $\tau = 0.50$ yielding $A = 0.975$. Clearly, if it is known that both misclassifications yield an equivalent "potentially large impact" or if nothing is known *a priori* about the misclassification costs, usage of Eq. 4 as a metric cannot be warranted.

However, when one does have prior knowledge that outcomes are economically imbalanced, a goal-oriented[28,30] approach that considers the economic costs is appropriate. For simplicity, one can define

$$\mathcal{C} = \frac{\mathcal{C}_{\mathsf{FN}}}{\mathcal{C}_{\mathsf{FP}}}, \mathcal{C}_{\mathsf{FN}} \gg \mathcal{C}_{\mathsf{FP}}. \tag{6}$$

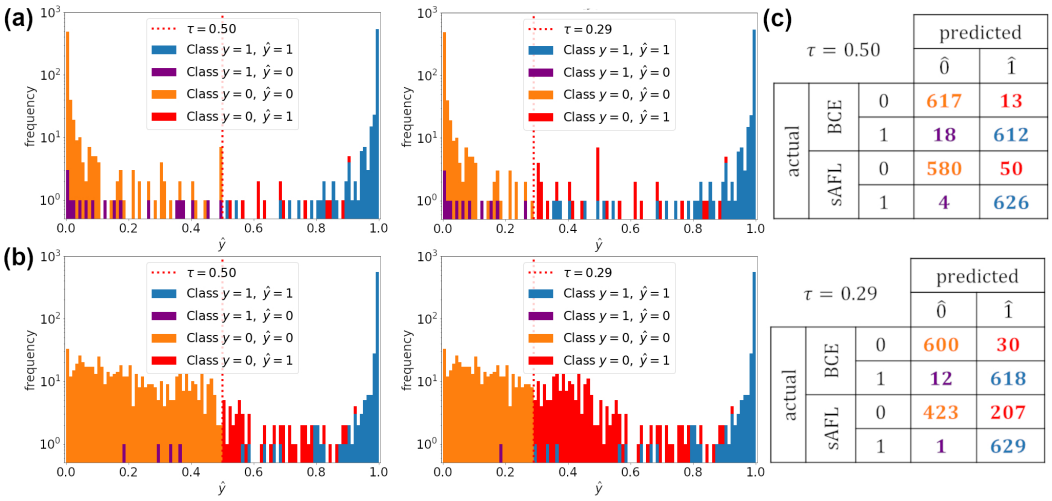

Figure 1. (a) Distributions of labels from a trained convolutional neural network (CNN) and the binary cross-entropy (BCE) loss function. (b) Distribution of labels from a CNN trained using a scaled asymmetric focal loss (sAFL). (c) Confusion matrices for the BCE, sAFL at two differnt classification thresholds τ.

Noting that the number of misclassifications is dependent upon the location of the classification threshold τ, Eq. 4 and 6 can be combined to simplify the total economic misclassification cost in this work,

$$\mathcal{R}(\mathcal{C}, \tau) = \mathsf{FP}(\tau) + \mathcal{C} \cdot \mathsf{FN}(\tau). \tag{7}$$

The cost ratio $\mathcal{C}$ remains quantitatively undefined, and the data cannot inform us of the actual cost associated with the misclassification. However, one may compare ML results to determine at which $\mathcal{C}$ would the total cost be minimized, given two classifications of the same input data. Taking the confusion matrices in Fig. 1(c) for $\tau = 0.29$, one can calculate the cost ratio for which it is more beneficial to use the CNN with an asymmetric loss function from Fig. 1(b) instead of the CNN with a symmetric loss function in Fig. 1(a). The cross-over point must satisfy

$$\mathcal{R}(\mathcal{C}, \tau = 0.29) = (12 + 30 \cdot \mathcal{C}) = (1 + 207 \cdot \mathcal{C}) \rightarrow \mathcal{C} \approx 16.1, \tag{8}$$

thus for $\mathcal{C} \geq 16.1$ the CNN using an asymmetric loss function yields less total cost for these test data for $\tau = 0.29$.

The symmetric loss function for Fig. 1(a) is the BCE as denoted in Fig. 1(c). The loss function for Fig. 1(b) is the "sAFL" which stands for a version of the AFL (Eq. 3) that is scaled in this work to be directly comparable on average to the BCE. The scaled asymmetric loss function (sAFL) is

$$\mathcal{L}_{\text{sAFL}} = -\left(\frac{1}{w_s(\gamma)}\right) y(1-\hat{y})^{\gamma} \log(\hat{y}) - (1-y)\log(1-\hat{y}), \tag{9}$$

where $w_s(\gamma)$ accounts for the difference in average weighting of the focal loss term compared to the log-loss term. It can be shown that

$$\int_0^1 -\log(\hat{y}) d\hat{y} = 1,$$

and

$$\int_0^1 -\hat{y}^{\gamma} \log(1-\hat{y}) d\hat{y} = \frac{\mathcal{H}_{1+\gamma}}{1+\gamma} = \frac{\Psi(2+\gamma) + \gamma_e}{1+\gamma} \equiv w_s \leq 1, \tag{10}$$

where $\mathcal{H}_{1+\gamma}$ is the $(1+\gamma)^{\text{th}}$ harmonic number, Ψ is the digamma function, and γ_e is the Euler-Mascheroni constant. This under-weighting scales with γ non-linearly. For $\gamma = 0$, $\mathcal{H}_1 = 1$ and the FL, AFL, and sAFL all revert to Eq. 2, the binary cross-entropy.

4. NUMERICAL EXPERIMENT

4.1 Methods

4.1.1 Customizing the loss functions in binary classification

We seek optimal loss functions and values for the hyperparameter γ that minimizes total costs $\mathcal{R}(\mathcal{C}, \tau)$. The numerical experiment compares five types of loss functions as illustrated in Fig. 2. Three are identified above:

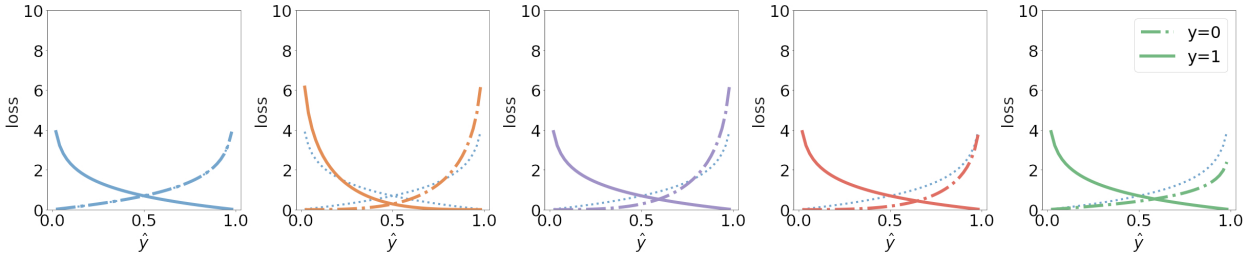

Figure 2. Five types of loss functions compared in this work, shown for $\gamma = 2.0$. The BCE is shown in all panels as thin dotted lines; the area below each function is the same for the BCE, sFL, and SAFL; the area below the functions for AFL are wBCE are equivalent to each other.

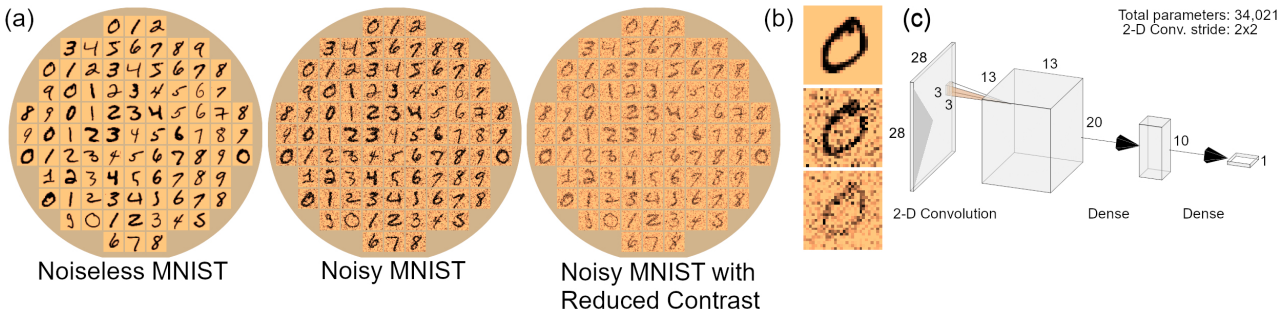

Figure 3. Validating ML concepts for defect metrology using publicly available ML data sets. (a) Three versions of the MNIST data set are used to represent three different measurement conditions. (b) Close-up view of noise and contrast differences among the sets. (c) Convolutional neural network (CNN) topography employed. Noiseless MNIST curated by LeCun *et al.*[34] Noisy MNIST data by Basu, *et al.*[35,36]

the BCE, the AFL, and the sAFL, with the sAFL and BCE having the same average magnitude across $\hat{y} \in [0, 1]$. It is straightforward to compare a third function against these, a scaled focal loss (sFL)

$$\mathcal{L}_{\text{sFL}} = -\left(\frac{1}{w_s(\gamma)}\right)\left(y(1-\hat{y})^{\gamma}\log(\hat{y}) + (1-y)\log(1-\hat{y})\right). \tag{11}$$

As the AFL however has intrinsic weighting, a fifth loss function is required for equally weighted comparisons to an established loss function, a weighted BCE (wBCE)

$$\mathcal{L}_{\text{wBCE}} = -w_s(\gamma)y\log(\hat{y}) - (1-y)\log(1-\hat{y}), \tag{12}$$

that scales non-linearly with γ.

4.1.2 Data sets and neural network

As the Authors have reported both simulated[19,31] and experimental[32,33] image data from defective and nominal semiconductor patterns previously, the limited number of images informed the decision to assess these loss functions using labeled ML image data sets. First, results reported should be wholly reproducible which requires open access image sets. Second, although data augmentation techniques have been previously employed for electromagnetically simulated images,[19] using these ML data sets without data augmentation reduces the possibility that the data might be biased leading to potentially inaccurate conclusions from the numerical experiment.

Three specific ML data sets have been used for the numerical experiment, each based upon the famous MNIST data set. For examining total misclassification cost in defect metrology, assume these data represent images from three different "measurement" conditions of the same underlying "wafer". Figure 3(a) shows 89 image results from one "wafer" out of around seventy wafers total that comprise our image sets. A ideal measurement condition yields patterns without noise, the "Noiseless MNIST" set which is the "MNIST-784" data set curated by LeCun.[34] Two more realistic measurements yield what is referred to in this work as the "Noisy MNIST" and "Noisy MNIST with reduced contrast" data set from Basu, et al.[35,36] The latter set has reduced the contrast of the underlying digits while both sets show added white Gaussian noise. One example of the variations among data sets are highlighted as Fig. 3(b).

Each data set contains 70000 images total of the ten handwritten digits $0, \ldots, 9$ however the digits are in unequal proportions. For reproducible results and a 1:1 ratio between classes for each digit pair, the random seed was fixed prior to a pseudo-random draw of 6300 examples per digit. This work synchronizes the ordering of the image set of Basu *et al.* to that of LeCun *et al.* allowing the random draw to yield image sets containing the same underlying digit for all "measurement" conditions. A 80 % training, 10 % validation, and 10 % test split was used for establishing the hyperparameters, yielding 630 test examples per digit in this paper. With this limited number of images, imposing class imbalance on the validation and test data was impractical; ramifications of this are discussed as Sec. 5.1.

Table 3. Values for hyperparameter γ and their effective weights ($w_s(\gamma)$) for the wBCE.

γ	0.4	0.8	1.2	1.6	2.0	3.0	4.0	5.0	6.0	8.0	10.0
$w_s(\gamma)$	0.8786	0.7877	0.7164	0.6588	0.6111	0.5208	0.4567	0.4083	0.3704	0.3143	0.2745

MNIST images of one digit are Class 0, the "nominal" patterning case while images of another digit are Class 1, the "defect" case. All 90 possible iterations among the digits have been compared, allowing each digit to be the "nominal" to every other digit's "defect" and *vice versa*. The CNN utilized is illustrated as Fig. 3(c). The numerical experiment has been performed using TensorFlow 2.9.0*[37] using custom loss functions with sigmoid activation at the final layer and ReLU elsewhere. Learning rates were 0.002 for the Noiseless MNIST data and 0.001 for both sets of noisy MNIST data using the Ftrl[38] optimizer that features built-in learning rate decay (using default decay power $= -0.5$). Each pair-wise training has been performed using ten epochs and a batch size of 100.

4.1.3 Design of experiment

Of the five loss function types in Fig. 2, four accept a hyperparameter, γ. As listed in Table 3, eleven values for γ have been used with the four loss functions; with one training instance with the BCE, 45 total loss functions have been utilized in the training of CNNs for each digit pair. The default value from the literature for the focal loss is $\gamma = 2.0$ while $\gamma > 5.0$ is reportedly problematic[26] but larger γ are included to test the (s)AFL and to allow greater weighting differences in the wBCE.

4.2 Minimizing total misclassification cost for each function

4.2.1 Misclassifications per function

Each panel of Fig. 4(a) shows 90 separate curves overlaid, one for each digit-pair, indicating the number of images misclassified as false positives FP and false negatives FN for the three "measurement" data sets across $\tau = 0.01, \ldots, 0.99$. Shown at $\gamma = 2.0$, the general trend is that for $\tau < 0.5$, there is a decrease in FN for the loss functions associated with the focal loss (sFL, sAFL, and AFL) relative to those associated with the binary cross-entropy (BCE, wBCE). The consequence of minimizing these FN is a dramatic increase in FP especially as $\tau \to 0$.

One implication from these data is that for some digit pairs, at small τ focal losses either nearly or exactly yield a "null classifier" (i.e. all images are predicted as Class 1, "defects"). A legitimate concern is that summing

*Certain commercial materials are identified in this paper in order to specify the experimental procedure adequately. such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials are necessarily the best available for the purpose.

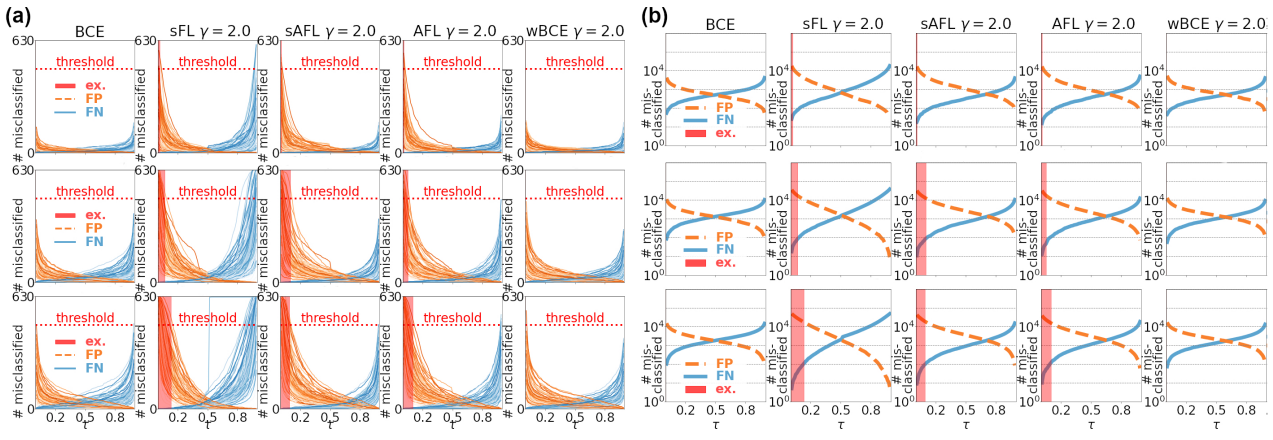

Figure 4. (a) Individual and (b) aggregated misclassifications for the three "measurement" data sets. The top row is from Noiseless MNIST; middle row, from Noisy MNIST; bottom row, Noisy MNIST with reduced contrast. Red boxes in the figure show combinations of loss function and τ excluded from consideration in minimizing total misclassification costs.

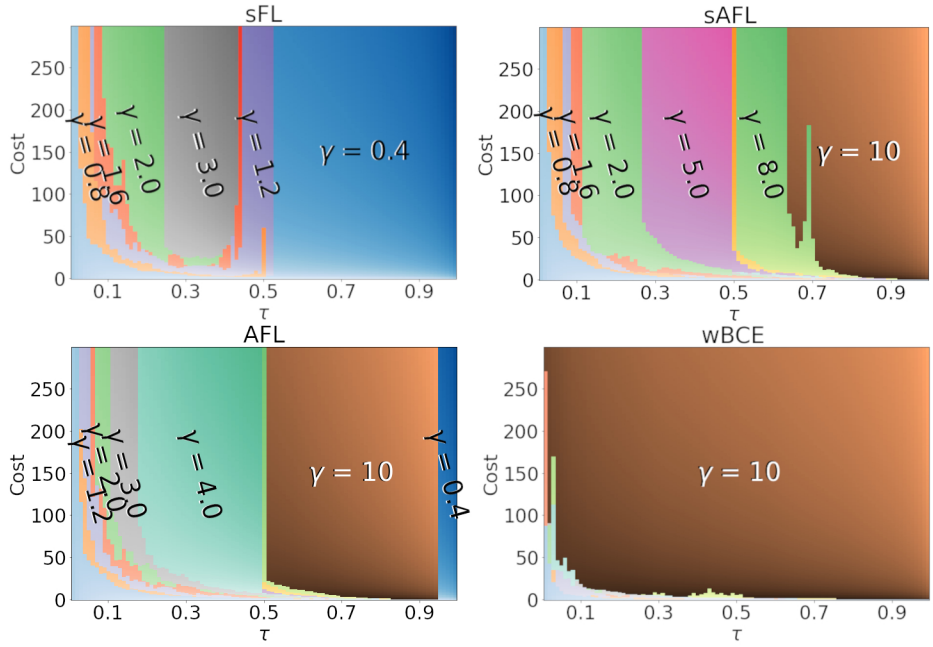

Figure 5. Maps of the best hyperparameter γ to minimize the total economic misclassification cost for four different loss functions as functions of cost ratio $\mathcal{C}$ and classification threshold τ. Note that the binary cross-entropy is independent of γ.

all misclassifications together in Fig. 4(b) hides poor classifiers and in a defect metrology setting one would not want to employ such a classifier to the hardest-to-compare examples. Therefore, a threshold has been added to Fig. 4(a) at $\mathsf{FP} = 470$, or approximately three-quarters misclassified, with red boxes showing the region of τ for which any of the 90 digit-pairs yield FP above the threshold. These regions are repeated in Fig. 4(b) to denote pairings of loss function $\mathcal{L}$ and τ that have been excluded from consideration as an optimal loss function for reducing the total cost.

4.2.2 Optimal hyperparameter per loss function

Figure 5 displays the best hyperparameter γ for reducing $\mathcal{R}(\mathcal{C}, \tau)$, the total cost as functions of cost and threshold, for four individual loss function types for the "Noisy MNIST" data set. For the weighted BCE, $\gamma = 10$ is the optimal solution over most of the parameter space shown. The sAFL and AFL trend similarly for $\tau \leq 0.5$ with best γ decreasing as τ decreases. As indicated by the relatively large number of FN, FN in Fig. 2(a) for the sFL, the symmetric focal loss is a poor choice for this application, and its distinct functional curvature does not improve the total cost when applied symmetrically.

4.3 Minimizing total economic misclassification cost across loss functions

Plotting the best γ for each loss function dependent upon γ yields the four plots in Fig. 5. Not shown here are the 11 plots of the best loss function for each γ, including the BCE. It is from these 11 plots that the best γ and loss function combinations at selected $\mathcal{C}$ have been extracted as functions of τ. Specific plots at all τ are shown as Fig. 6 for $\mathcal{C} = 16, 128, 1024$. These are augmented with data for specific τ in Table 4 using $\mathcal{C} = 2^n,\ n = 0, \ldots, 10$. The numerical values for r_{red} in Table 4 are

$$\mathsf{r}_{\text{red}}(\mathcal{C}, \tau) = 1 - \frac{\mathcal{R}_{\mathcal{L}_{\text{best}}}(\mathcal{C}, \tau)}{\mathcal{R}_{\mathcal{L}_{\text{wBCE}}}(\mathcal{C}, \tau)}, \tag{13}$$

where $\mathcal{R}_{\mathcal{L}_{\text{best}}}$ is the smallest total cost for $\mathcal{C}, \tau$ using the best loss function $\mathcal{L}_{\text{best}}$, and $\mathcal{R}_{\mathcal{L}_{\text{wBCE}}}$ is the smallest total cost for $\mathcal{C}, \tau$ using a valid weighted binary-cross entropy loss $\mathcal{L}_{\text{wBCE}}$. For example, for $\mathcal{C} = 64, \tau = 0.3, \mathcal{L}_{\text{best}} = \mathcal{L}_{\text{AFL}}(\gamma = 5)$ which yields at least a 19 % improvement for these data sets over the optimal wBCE result. It is

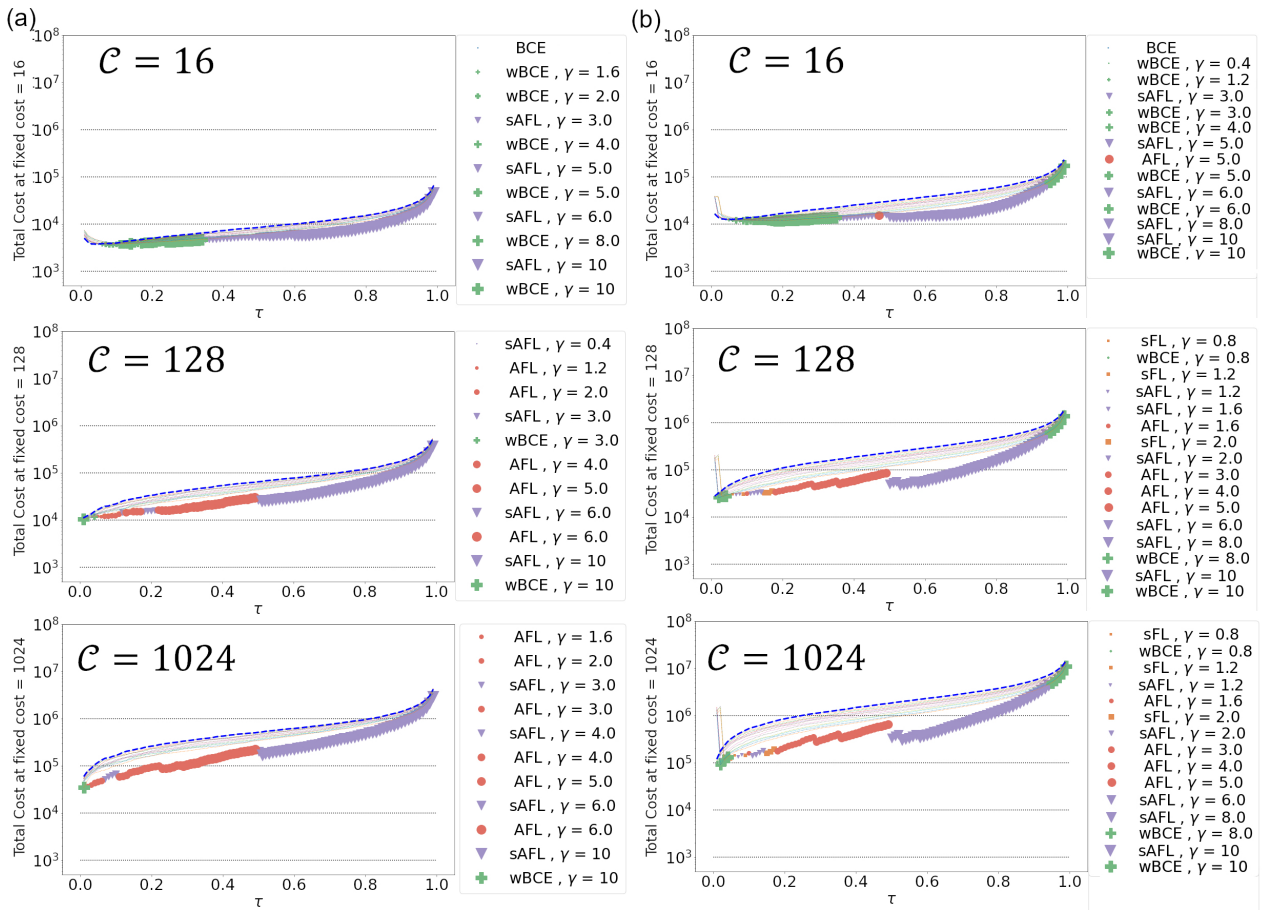

Figure 6. Total economic misclassification cost (total cost) for the (a) Noiseless MNIST and (b) Noisy MNIST with reduced contrast data sets for three values of $\mathcal{C}$. Blue dashed line represents the BCE result, thinner lines represent results from the wBCE using Table 3. Size of the marker scales with γ.

"at least" a 19 % improvement as the worst result among the three data sets is shown in Table 4 to establish a lower bound for these data and stated experimental conditions. Such caution is necessary as a single, fixed random seed is employed to obtain reproducible results.

There are key trends observed for three loss functions from this numerical experiment. First, for $\tau \approx 0.5$ and $\mathcal{C} \geq 16$, one should use the sAFL and $\gamma = 5$ or $\gamma = 6$ for these data sets to minimize the total economic misclassification cost. This suggests that the asymmetric difference in curvature due to Eq. 9 is more important than weighting effects from Table 3. The sAFL is only recommended for $\mathcal{C} < 64, 0.2 \leq \tau \leq 0.5$ for these data. Second, the use of the AFL is in general suggested for $0.1 \leq \tau < 0.5$ for $\mathcal{C} \geq 64$. This trend is observable in the middle and bottom rows of Fig. 6. There is a general decrease in optimal γ as τ decreases. As $\mathcal{C}$ increases, both the intrinsic weighting and the difference in curvature are important for minimizing $\mathcal{R}$; if weighting alone accounted for this increase, then the wBCE should instead dominate. Third, as $\tau \to 0$, especially for reduced $\mathcal{C}$, the wBCE does outperform the asymmetric loss functions tested here. Although asymmetric focal losses do not reduce the total economic misclassification cost for all possible $\mathcal{C}$ and τ, these results using a 1:1 ratio between "defect"-indicating and "nominal"-indicating data in training, test, and validation indicate that tailoring the loss function may be a viable method for manipulating the total economic misclassification cost especially for applications with high misclassification cost ratios such as defect metrology.

Table 4. Best loss function $\mathcal{L}(\gamma)$ and percent reduction $\mathsf{r}_{\text{red}}(\%)$ compared to $\mathcal{L}_{\text{wBCE}}$. Due to space limitations, function names are further abbreviated: "sA" = sAFL, "AF" = AFL, "wB" = wBCE, "sF" = sFL.

	τ											
	0.5		0.4		0.3		0.2		0.1		0.05	
$\mathcal{C}$	$\mathcal{L}(\gamma)$	r_{red}	$\mathcal{L}(\gamma)$	r_{red}	$\mathcal{L}(\gamma)$	r_{red}	$\mathcal{L}(\gamma)$	r_{red}	$\mathcal{L}(\gamma)$	r_{red}	$\mathcal{L}(\gamma)$	r_{red}
16	sA,5.0	14	sA,3.0	5	wB,10	n/a	wB,8.0	n/a	wB,2.0	n/a	BCE	n/a
32	sA,6.0	25	sA,3.0	14	sA,3.0	8	AF,2.0	1	wB,10	n/a	wB,3.0	n/a
64	sA,6.0	33	AF,6.0	24	AF,5.0	20	sA,3.0	15	AF,1.2	3	wB,3.0	n/a
128	sA,6.0	37	AF,6.0	33	AF,6.0	33	sA,3.0	28	AF,2.0	16	AF,1.2	1
256	sA,6.0	39	AF,6.0	37	AF,6.0	42	sA,3.0	34	AF,2.0	25	AF,2.0	16
512	sA,6.0	40	AF,6.0	39	AF,6.0	47	AF,4.0	39	AF,2.0	30	AF,2.0	28
1024	sA,6.0	41	AF,6.0	40	AF,6.0	50	AF,4.0	42	sA,4.0	34	AF,3.0	37

5. APPLICABILITY OF THIS WORK TO DEFECT INSPECTION

5.1 Class imbalance in defect metrology

The primary challenge to the industrial applicability of the results of the numerical experiment is the 1:1 ratio between the "defect" and "nominal" classes in test and validation. In industrial application with binary classification, the economic imbalance $\mathcal{C}$ may be comparable to a class imbalance with most data indicating "nominal" patterning. To address this, one can define

$$\mathcal{M} = \frac{\mathsf{N}}{\mathsf{P}}, \tag{14}$$

where for the moment the "defect" class is Class 1 and "Positive" with P members, and the "nominal" class is Class 0 and "Negative" with N members. If indeed $\mathsf{N} \gg \mathsf{P}$ then the results in Sec. 4.1.3 must be viewed as a *normalized* total economic misclassification cost. One may *estimate* using $\mathcal{M}$ that

$$\mathsf{N} \approx \mathcal{M} \cdot \mathsf{FP}_{\mathsf{study}} + \mathcal{M} \cdot \mathsf{TN}_{\mathsf{study}}, \tag{15}$$

where $\mathsf{FP}_{\mathsf{study}}$ and $\mathsf{TN}_{\mathsf{study}}$ represent this numerical experiment's classifications as expressed in Table 2. One can restate the total economic misclassification cost then as

$$\mathcal{R} \approx \mathcal{M} \cdot \mathsf{FP}_{\mathsf{study}} + \mathcal{C} \cdot \mathsf{FN}_{\mathsf{study}}, \tag{16}$$

where $\mathsf{FN}_{\mathsf{study}}$ is also defined from the numerical experiment. If $\mathcal{C} > \mathcal{M}$, then the normalized total cost is

$$\acute{\mathcal{R}} \approx \mathsf{FP}_{\mathsf{study}} + \acute{\mathcal{C}} \cdot \mathsf{FN}_{\mathsf{study}}, \acute{\mathcal{C}} = \frac{\mathcal{C}}{\mathcal{M}}, \mathcal{C} > \mathcal{M}. \tag{17}$$

However, if $\mathcal{M} > \mathcal{C}$, then Eq.17 cannot be considered. In these conditions, the presupposition undergirding Eqs. 3 and 9 is no longer valid - if $\mathcal{M} > \mathcal{C}$, the defect Class 1 is qualitatively *less* important to the total normalized economic misclassification cost than the nominal Class 0. Assume then that the "nominal" class is Class 1, 'positive' with P member and the "defect" class is Class 0, 'negative' with N members. It can be shown that a different form for the normalized total cost is

$$\breve{\mathcal{R}} \approx \mathsf{FP}_{\mathsf{study}} + \breve{\mathcal{C}} \cdot \mathsf{FN}_{\mathsf{study}}, \breve{\mathcal{C}} = \frac{\mathcal{M}}{\mathcal{C}}, \mathcal{M} > \mathcal{C}. \tag{18}$$

The term $\acute{\mathcal{C}}$ or $\breve{\mathcal{C}}$ is substituted for $\mathcal{C}$ in the results in Sec. 4. Clearly defining which of the two total normalized misclassification costs $(\acute{\mathcal{R}}, \breve{\mathcal{R}})$ is being referenced becomes essential.

5.2 Other considerations

There are more challenges to extending the lessons learned through this numerical experiment to industrial semiconductor inspection. For example, if a semiconductor manufacturer has an unvarying, accurate value for $\mathcal{C}$, then optimization of a goal-based metric would likely not require a comparison among multiple asymmetric loss functions and their hyperparameters. If $\mathcal{C}$ is known to fluctuate, then a similar study to map out ideal conditions

may be indicated. A multi-class defect environment, where "nominal" is one class among many (e.g., "bridge", "void", etc.) may arise in SEM defect metrology. It is straightforward to re-write a more general loss function $\mathcal{L}_{\text{gen}}$ as

$$\mathcal{L}_{\text{gen}} = -\sum_{i=1}^{m} y_i (1-\hat{y})^{\gamma_i} \log(\hat{y}_i), \tag{19}$$

where m is the number of classes, y_i and $\hat{y}_i, i = 1, \ldots, m$ are the actual and predicted labels respectively, and γ_i is the hyperparameter for each class. For emphasizing one class (e.g., Class 4 of 5) it may be sufficient to vary a scalar γ where $\boldsymbol{\gamma} = [0, 0, 0, \gamma, 0]$. Setting $\gamma = 0$ would yield the categorical cross-entropy. Such optimization has not been performed here as optical defect inspection tools can at best only identify the *presence* of a defect and not its *type*, thus binary classification is more appropriate as a starting point for this most challenging manufacturing problem.

6. CONCLUSIONS

Misclassification costs per class for industrial applications such as semiconductor patterned defect inspection may vary greatly between the "defect" class(es) and the "nominal" class, with much greater economic consequences for misclassifying a "defect" as "nominal". A goal-oriented approach to minimize the total economic misclassification cost (total cost) has been pursued with five types of loss functions used in training CNNs, two types of which are asymmetric focal losses (AFLs). For three related publicly available data sets, a scaled version of the AFL decreased the total cost for instances where the misclassification cost ratio $\mathcal{C} \geq 16$ at a classification threshold $\tau = 0.5$, the default position for most binary classifiers. Shifting this threshold to $0.1 \leq \tau < 0.5$ yields lower total costs if the cost ratio is $\mathcal{C} > 128$. Overall, the asymmetric loss functions lower total cost by 15 % to 40 % for $0.2 < \tau < 0.5$ for $\mathcal{C} > 64$ compared to comparably weighted binary cross-entropies. Such large values of the cost ratio $\mathcal{C}$ should be expected not only from semiconductor defect metrology but also from many other applications.

REFERENCES

[1] Franke, J.-H., Frommhold, A., Dauendorffer, A., Nafus, K., Rispens, G., and Maslow, M., "Elucidating the role of imaging metrics for variability and after etch defectivity," *Journal of Micro/Nanopatterning, Materials, and Metrology* **21**(2), 023201–023201 (2022).

[2] Meshulach, D., Dolev, I., Yamazaki, Y., Tsuchiya, K., Kaneko, M., Yoshino, K., and Fujii, T., "Advanced lithography: wafer defect scattering analysis at DUV," *Proc. SPIE* **7638**, 195–204 (2010).

[3] Patterson, O. D., Lee, J., Salvador, D. M., Lei, S.-C. C., and Tang, X., "Detection of sub-design rule physical defects using e-beam inspection," *IEEE Trans. Semicond. Manuf.* **26**(4), 476–481 (2013).

[4] Sarkar, S. K., Das, S., Carballo, V. M. B., Leray, P., and Halder, S., "Investigating metal oxide resists for patterning 28-nm pitch structures using single exposure extreme ultraviolet: defectivity, electrical test, and voltage contrast study," *Journal of Micro/Nanopatterning, Materials, and Metrology* **21**(4), 044901 (2022).

[5] Li, Z. J., Kamnitsas, K., and Glocker, B., "Overfitting of neural nets under class imbalance: Analysis and improvements for segmentation," in [*10th International Workshop on Machine Learning in Medical Imaging (MLMI) / 22nd International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*], *Lecture Notes in Computer Science* **11766**, 402–410 (2019).

[6] Imoto, K., Mishima, S., Arai, Y., and Kondo, R., "Impact of data imbalance caused by inactive frames and difference in sound duration on sound event detection performance," *Applied Acoustics* **196**, 108882 (2022).

[7] Vogt, N., Arya, Z., Nunez, L., Hobson, K., Connell, J., Brady, S. M., and Aljabar, P., "A deep-learning lesion segmentation model that addresses class imbalance and expected low probability tissue abnormalities in pre and postoperative liver MRI," in [*26th Annual Conference on Medical Image Understanding and Analysis (MIUA)*], *Lecture Notes in Computer Science* **13413**, 398–411 (2022).

[8] Chen, B., Zhang, L., Liu, T. T., Li, H. S., and He, C., "Lightweight network with variable asymmetric rebalancing strategy for small and imbalanced fault diagnosis," *Machines* **10**(10), 879 (2022).

[9] Tran, T. K., "Defect characterization and metrology," in [*Metrology and Diagnostic Techniques for Nanoelectronics*], Ma, Z. and Seiler, D. G., eds., 589–635, Pan Stanford, Singapore (2017).

[10] Lee, D.-H., Yang, J.-K., Lee, C.-H., and Kim, K.-J., "A data-driven approach to selection of critical process steps in the semiconductor manufacturing process considering missing and imbalanced data," *Journal of Manufacturing Systems* **52**, 146–156 (2019).

[11] Hess, C. and Weiland, L. H., "Issues on the size and outline of killer defects and their influence on yield modeling," in [*IEEE/SEMI 1996 Advanced Semiconductor Manufacturing Conference and Workshop*], *IEEE/SEMI 1996 Advanced Semiconductor Manufacturing Conference and Workshop*, 423–428 (1996).

[12] LaPedus, M., "Finding defects is getting harder." *Semiconductor Engineering*, http://semiengineering.com/finding-killer-defects/ (2016). Accessed 30 Apr 2018.

[13] Zhu, J., Liu, J., Xu, T., Yuan, S., Zhang, Z., Jiang, H., Gu, H., Zhou, R., and Liu, S., "Optical wafer defect inspection at the 10 nm technology node and beyond," *International Journal of Extreme Manufacturing* (2022).

[14] "International roadmap for devices and systems (IRDS™) 2021 edition: Metrology." https://irds.ieee.org/editions/2021/metrology (2021). Accessed 16 Feb 2022.

[15] Bischoff, J., Bauer, J., Haak, U., Hutschenreuther, L., and Truckenbrodt, H., "Optical scatterometry of quarter micron patterns using neural regression," *Proc. SPIE* **3332**, 526–537 (1998).

[16] Jeon, K. A., Kim, H. H., Yoo, J. Y., Park, J. T., and Oh, H. K., "The extraction of exposure parameters by using neural networks," *Proc. SPIE* **5039**, 1105–1114 (2003).

[17] Kuo, H. F. and Faricha, A., "Artificial neural network for diffraction based overlay measurement," *IEEE Access* **4**, 7479–7486 (2016).

[18] Kvamme, H., Sellereite, N., Aas, K., and Sjursen, S., "Predicting mortgage default using convolutional neural networks," *Expert Systems with Applications* **102**, 207–217 (2018).

[19] Henn, M.-A., Zhou, H., and Barnes, B. M., "Data-driven approaches to optical patterned defect detection," *OSA Continuum* **2**(9), 2683–2693 (2019).

[20] Dey, B., Dehaerne, E., Halder, S., Leray, P., and Bayoumi, M. A., "Deep learning based defect classification and detection in SEM images: a mask R-CNN approach," *Proc. SPIE* **12053**, 120530K (2022).

[21] Devika, B. and George, N., "Convolutional neural network for semiconductor wafer defect detection," in [*2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*], 1–6 (2019).

[22] Ahn, J., Kim, Y. C., Kim, S. Y., Hur, S.-M., and Thapar, V., "Defect recognition in line-space patterns aided by deep learning with data augmentation," *Journal of Micro/Nanopatterning, Materials, and Metrology* **20**(4), 041203 (2021).

[23] Fukuda, K., Ouchi, M., Ishikawa, M., Yoshida, Y., Fukaya, K., Kagetani, R., and Shindo, H., "Trainable die-to-database for large field of view e-beam inspection," *Journal of Micro/Nanopatterning, Materials, and Metrology* **22**(2), 021004 (2022).

[24] Fukuda, H. and Kondo, T., "Anomaly detection in random circuit patterns using autoencoder," *Journal of Micro/Nanopatterning, Materials, and Metrology* **20**(4), 044001 (2021).

[25] Johnson, J. M. and Khoshgoftaar, T. M., "Survey on deep learning with class imbalance," *Journal of Big Data* **6**(1), 1–54 (2019).

[26] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P., "Focal loss for dense object detection," in [*2017 IEEE International Conference on Computer Vision (ICCV)*], 2999–3007 (2017).

[27] Ho, Y. and Wookey, S., "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling," *IEEE Access* **8**, 4806–4813 (2019).

[28] Draper, B. A., Brodley, C. E., and Utgoff, P. E., "Goal-directed classification using linear machine decision trees," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **16**(9), 888–893 (1994).

[29] Gerunov, A., "Binary classification problems in economics and 136 different ways to solve them," *Bulgarian Economic Papers* (2), 2–31 (2020).

[30] Yan, W. Z., Goebel, K., and Li, J. C., "Classifier performance measures in multi-fault diagnosis for aircraft engines," *Proc. SPIE* **4733**, 88–97 (2002).

[31] Barnes, B. M., Henn, M.-A., Sohn, M. Y., Zhou, H., and Silver, R. M., "Assessing form-dependent optical scattering at vacuum- and extreme-ultraviolet wavelengths of nanostructures with two-dimensional periodicity," *Physical Review Applied* **11**(6), 064056 (2019).

[32] Barnes, B. M., Sohn, M. Y., Goasmat, F., Zhou, H., Vladar, A. E., Silver, R. M., and Arceo, A., "Three-dimensional deep sub-wavelength defect detection using λ=193 nm optical microscopy," *Optics Express* **21**(22), 26219–26226 (2013).

[33] Barnes, B. M., Goasmat, F., Sohn, M. Y., Zhou, H., Vladar, A. E., and Silver, R. M., "Effects of wafer noise on the detection of 20-nm defects using optical volumetric inspection," *Journal of Micro-Nanolithography MEMS and MOEMS* **14**(1) (2015).

[34] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P., "Gradient-based learning applied to document recognition," *Proceedings of the IEEE* **86**(11), 2278–2324 (1998).

[35] Basu, S., Karki, M., DiBiano, R., Mukhopadhyay, S., Ganguly, S., and Nemani, R. R., "The N-MNIST handwritten digit dataset." https://csc.lsu.edu/~saikat/n-mnist/ (2014). Accessed 4 Dec 2020.

[36] Basu, S., Karki, M., Ganguly, S., DiBiano, R., Mukhopadhyay, S., Gayaka, S., Kannan, R., and Nemani, R., "Learning sparse feature representations using probabilistic quadtrees and deep belief nets," *Neural Processing Letters* **45**, 855–867 (2017).

[37] Abadi, M., Agarwal, A., Barham, P., et al., "TensorFlow: Large-scale machine learning on heterogeneous systems," (2015). Software available from tensorflow.org.

[38] McMahan, H. B., Holt, G., Sculley, D., Young, M., Ebner, D., Grady, J., Nie, L., Phillips, T., Davydov, E., Golovin, D., Chikkerur, S., Liu, D., Wattenberg, M., Hrafnkelsson, A. M., Boulos, T., and Kubica, J., "Ad click prediction: A view from the trenches," in [*Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*], *KDD '13*, 1222–1230, Association for Computing Machinery, New York, NY, USA (2013).