An IEEE Societies and Technical Councils Publication
IEEE Data Descriptions

Received 10 January 2025; revised 14 April 2025, 16 April 2025, and 16 April 2025; accepted 16 April 2025; Date of publication 21 April 2025; date of current version 30 May 2025.

Digital Object Identifier 10.1109/IEEEDATA.2025.3562805

Descriptor: *Not-A-DAtabase of Synthetic Shapes Benchmarking Dataset (NADA-SynShapes)*

GIULIO DEL CORSO, FEDERICO VOLPINI, CLAUDIA CAUDAI, DAVIDE MORONI, AND SARA COLANTONIO

Institute of Information Science and Technologies (ISTI), National Research Council of Italy (CNR) - Pisa, 56124 Pisa, Italy

CORRESPONDING AUTHOR: Claudia Caudai (e-mail: claudia.caudai@isti.cnr.it).

This work was supported in part by the European Union's Horizon 2020 Research and Innovation Program (*ProCAncer-I*) under Grant 952159 and in part by the European Union's Horizon Europe Research and Innovation Program (*FAITH*) under Grant 101135932.
(Giulio Del Corso, Federico Volpini, and Claudia Caudai contributed equally to this work.)

ABSTRACT NADA is a synthetic dataset of 1 500 000 (128 × 128 pixels) images of geometric shapes with non-elementary multivariate parameter distributions designed to benchmark and test novel probabilistic deep learning models. Benchmarking uncertainty-aware techniques is critical in real-world scenarios, especially in high-stakes domains such as automated driving or health data analysis. This is because the most robust and reliable AI methods are rooted in Bayesian reasoning and uncertainty analysis, yet there is no consensus on how to test or compare them. In particular, the public, synthetic, and real datasets currently available in the literature are inadequate for benchmarking most probabilistic methodologies: first, due to a lack of control over population variability; and second, due to an oversimplification of the distributional properties of latent variables. From this perspective, NADA is organized into three main From this perspective, NADA is organized into three main repositories, specifically developed to challenge uncertainty-aware methods and provide a unified benchmark reference dataset across three key areas: 1) characterization of complex latent space and evaluation of disentangling ability (NADA_Dis: 300 000 images); 2) identification of different types of aleatoric and epistemic uncertainties (NADA_AlEp: 500 000 images); and 3) detection of out-of-distribution elements for reliable AI (NADA_OOD: 700 000 images). Each repository includes the dataframe describing the image parameters (e.g., rotation, position, shape, color, deformation, and noise), the dataset-generating hyperparameters (e.g., marginal distributions and correlation matrix), and evaluation plots to assess data quality. In addition, the dataset is coupled with an open-source Python synthetic generator, allowing easy modification and adaptation to specific research questions.

IEEE SOCIETY/COUNCIL IEEE Reliability Society

DATA TYPE/LOCATION Images; Geometrical Shapes; Synthetic Data

DATA DOI/PID 10.5281/zenodo.14361220

INDEX TERMS Feature disentanglement, out-of-distribution identification, synthetic data generator, uncertainty quantification benchmarking.

BACKGROUND

The emerging popularity of computational models based on deep neural networks poses a challenge to researchers in terms of efficient training and standardized performance testing. With the growing use of deep learning as decision support in areas closely related to our daily lives, health, and safety, the need to test the reliability, reproducibility, and level of trustworthiness of these models is also growing, especially with respect to the need to estimate the uncertainty of predictions in several fields such as medicine, biology,

automation, and bioengineering [1], [2], [3], [4], [5], [6], [7]. In particular, probabilistic neural networks based on Bayesian mechanisms (generative networks, Bayesian neural networks [8], [9], etc.) are at the cutting edge of current scientific research but require specific and appropriate databases to be tested and analyzed. These difficulties are further exacerbated when the networks have to deal with images, both for classification and for regression problems, as there are even fewer standardized databases available. Indeed, despite the availability of numerous databases in the literature, most are not explicitly designed to be used to determine the predictive power of *probabilistic* models or to estimate the *reliability* of models capable of producing reliability estimates.

Public databases based on real data collections are closer to the application reality of a model but do not provide detailed information on the distributions of the latent variables that characterize images. In addition, while complex and natural images can provide a more effective benchmark for case-specific techniques, the development of foundational core methodologies must be rooted in simpler but effective test data. ImageNet [10] must certainly be included among the most popular examples. It is a large and freely available visual database designed for use in visual object recognition, containing more than 14 million hand-annotated images divided into more than 20 000 categories. The Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) dataset [11] consists of hours of traffic scenarios recorded from a moving platform with various sensor modalities, including camera images, laser scans, and high-precision GPS measurements. The natural images dataset [12] contains 6899 freely available images from eight distinct classes (airplane, car, cat, dog, flower, fruit, motorbike, and person). Among the natural image datasets, one of the most famous and used is Modified National Institute of Standards and Technology (MNIST) database [13], a dataset of size-normalized and centered handwritten digits available from National Institute of Standards and Technology (NIST), which has become a standard benchmark for learning, classification, and computer vision systems, consisting of a training set of 60 000 examples, and a test set of 10 000 examples. Recently, the extended MNIST (EMNIST) dataset [14], made of digits and letters, and the fashion MNIST dataset [15], made of Zalando's article images, have been introduced, for more challenging classification tasks.

Conversely, most of the available synthetic datasets usually provide very simplistic distributions that do not allow a detailed investigation of the potential of probabilistic networks to detect anomalies, outliers, or latent spaces of non-uniform and independent variables. Indeed, most synthetic datasets are created to test the disentanglement properties of unsupervised models. Among them, dSprites [16] is a popular dataset of 737 280 2-D shapes generated by combining exactly once six ground truth independent features (color, shape, scale, rotation, and x and y positions). The 2-D geometric shapes dataset [17] consists of nine classes of 2-D geometric shapes (triangle, square, pentagon, hexagon, heptagon, octagon, nonagon, circle, and star), each one made of 10k RGB images generated by selecting randomly and independently the perimeter, the position, the rotation, the background, and the filling color. The XYSquares dataset [18] is specifically constructed to test the disentanglement capabilities of VAE frameworks that use pixel-wise losses. It is built from three non-overlapping red, green, and blue 8×8 pixels squares, each moving along the x and y coordinates of a 64×64 pixel grid. Finally, the synthetic shapes dataset [19] comprises 600 high-quality images (circles, squares, and triangles) devoted to shape recognition tasks. However, these datasets have limitations, for example, the samples are taken only from uniform distributions, and therefore, they do not give the possibility of carrying out systematic analyses related to the specific characteristics of the datasets.

Therefore, even though users often rely on public datasets of natural images to study uncertainty in specific tasks (e.g., the LIDC dataset is used to evaluate uncertainty in U-Nets [20], and Mimori et al. [21] proposed an evaluation framework for class probability estimates in the presence of label uncertainty for DL models using a public medical image dataset [22]), it is also common practice to define ad hoc synthetic data tailored to the specific task [21], [23]. All the aforementioned strategies, however, lack standardization and user control due to noise in real data and task overspecialization in synthetic datasets and are therefore not appropriate for the systematic analysis and testing of foundational uncertainty-aware techniques.

In this article, we present *NADA* (Not-A-DAtabase), an open-source datasets generator, which is able to create in a few minutes entire complex datasets from scratch. In addition, we present three synthetic datasets (NADA-SynShapes), generated with NA_DA, tailored to answer three distinct questions in the testing and development of novel ML methodologies.

1) *[NADA_Dis: 300 000 Images]: Is the model able to correctly characterize a complex latent space?*
2) *[NADA_OOD: 700 000 Images]: Does the model detect out-of-distribution images?*
3) *[NADA_AlEp: 500 000 Images]: Does the model distinguish between different types of uncertainties?*

The datasets provided are synthetically generated through the use of a Python package, which we have developed and released under *Attribution-NonCommercial-NoDerivs 4.0 International* License (NA_DAtabase.git). The different types of uncertainties (reducible/irreducible) are modeled using non-trivial multivariate distributions to match the specific research questions. The collected data are then evaluated using auxiliary test functions that ensure correct data generation and display the provided empirical distributions.

Each synthetic dataset therefore includes the generated images, a dataframe describing the generation parameters, the dataframe containing all dataset properties to ensure the reproducibility of the generated images, and a folder containing the plots to evaluate the data quality.

COLLECTION METHODS AND DESIGN

The data provided are generated using a synthetic image generator that we developed to study the performance of probabilistic machine learning models [24]. The images were generated using multithreaded parallelization on an AMD Ryzen 7 5800H (3.2 GHz, 8 cores, 16 threads) CPU with 16 GB of RAM, taking approximately 3 h per folder. For each research question (NADA_Dis, NADA_OOD, and NADA_AlEp), we select the dataset parameters to define the required dataset characteristics.

Define Dataset Hyperparameters

The dataset hyperparameters are given as follows.

1) *Dataset Size*: Number of images generated.
2) *Sampling Strategy*: Technique for sampling the multivariate probability distribution. Possible values are Monte Carlo, Latin hypercube, and (Sobol') low discrepancy sequence.
3) *Random Seed*: Random seed can be specified to ensure full reproducibility of data generation.
4) *Resolution* (128 $\times$ 128): The resolution in pixels of the dimensions of the image (the canvas is a square).
5) *Background Color*: The color in hexadecimal value of the background of the image. *Allow out of border*: If TRUE, shapes can extend beyond the image boundaries.
6) *Mark Rotation*: If TRUE, the shapes are marked with an inner hole corresponding to the rotation value.

Class imbalance can be obtained by arbitrarily specifying the probability of each individual shape and color combination.

Define Images Marginal Distribution

As regards the characteristics of the individual images, the features that can be managed (described in Fig. 1) are color, shape, radius, rotation, position (x and y), and deformation. Deformation is a special uncertainty latent variable that acts on the image to change it from the original shape (i.e., an arbitrary regular polygon) to the circle. This is directly related to the ability of the model to correctly identify deformed shapes as unreliable predictions.

Introduce Aleatoric/Epistemic Uncertainties

Other uncertainties that affect the model include classification/labeling noise (quantifies the uncertainty in predicting a class for a given instance), blur (Gaussian), white noise (Gaussian), and additive/multiplicative noise (affects the regression values in an additive/multiplicative way). The aforementioned uncertainties can be summarized by the equation

$$\begin{aligned}\mathcal{M}_{\hat{\theta},\hat{h}}(\mathbf{x}) &\approx \mathcal{R}(\mathbf{x}) + \varepsilon_A + \varepsilon_E \\ &= \mathcal{R}(\mathbf{x}) + (\varepsilon_{A,I} + \varepsilon_{A,N}) + (\varepsilon_{E,M} + \varepsilon_{E,Ap}) \quad (1)\end{aligned}$$

where $\mathcal{M}_{\hat{\theta},\hat{h}}$ is the model, which depends on trainable parameters ($\hat{\theta}$) and non-trainable hyperparameters ($\hat{h}$), $\mathbf{x}$ is the input value (e.g., images in this dataset), $\mathcal{R}$ is the real relationship the model is trying to approximate, and $\varepsilon_A, \varepsilon_E$ are the aleatoric/epistemic uncertainties that pollute the results [25]. The aleatoric uncertainty ε_A is also called irreducible uncertainty because it does not depend on the amount of data available, and increasing the size of the training set does not shrink it. Such aleatoric uncertainty ε_A can be further subdivided into aleatoric experimental uncertainty/noise $\varepsilon_{A,N}$ and aleatoric inherent uncertainty $\varepsilon_{A,I}$. The first is described as labeling noise/regression noise, which randomly contaminates the results and limits predictive power. A more comprehensive and rigorous discussion of the different types of uncertainty can be found in the review by Del Corso et al. [25], while implementations of several uncertainty-aware techniques are freely available in the GitHub repository ShedLight_UQ. The dataset *NADA_AlEp_3_Label* incorporates $\varepsilon_{A,N}$. The second depends on the randomness contained in the real relationship $\mathcal{R}$ and is not modeled in this dataset.

The epistemic uncertainty ε_E, on the other side, stands for the reducible uncertainties, which can be reduced by increasing the amount of available data. It can be further subdivided into epistemic model uncertainty $\varepsilon_{E,M}$ (the difference between the optimal model and the relation) and epistemic approximation uncertainty $\varepsilon_{E,Ap}$ (the difference between the optimal model and the empirical one obtained due to a finite number of data). Several reducible amount of uncertainties are introduced in the following datasets, including deformation (*NADA_AlEp_2_Deformation*) or white noise (*NADA_AlEp_1_White_Noise*).

Define the multivariate distributions

The Dataset Generator allows detailed control over the multivariate distributions of the generation parameters and, thus, the latent space. In particular, using the Gaussian copula formalism and Sklar's theorem [26], a multivariate distribution can be uniquely characterized by its marginal distributions and a positive definite matrix representing the monotonic correlations among the marginals. Therefore, each dataset is generated by defining each marginal and then introducing their correlation to obtain a multivariate distribution. Continuous marginal distributions are selected from the maximum entropy distributions given their minimum and maximum values, their mean, and their standard deviation. Therefore, according to Shannon's entropy theory, they can be selected from uniform, Gaussian, and truncated Gaussian distributions [27]. The implementation of the multivariate distributions is based on classical uncertainty quantification

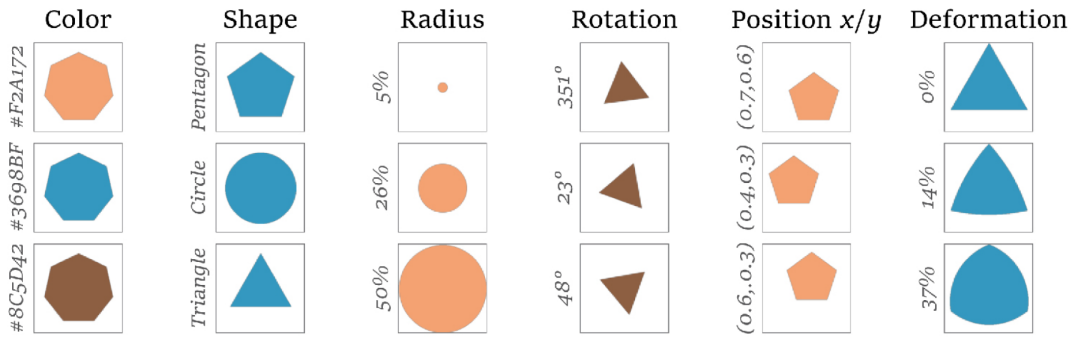

FIG. 1. **Image generation user-defined parameters: color (described by hex color codes), shape (arbitrary regular shapes given the number of vertices), radius (from 0% to 50% of the image), rotation (0°–360°), asymmetry can be induced in shape by enabling a corresponding option), position (center position in percentage from the top left of the image), and deformation (0%–100%, from the original regular shape to the circumscribing circle).**

libraries (OpenTURNS [28]) and is detailed in the technical report of the referenced GitHub repository [24].

VALIDATION AND QUALITY

The repositories provided in this article are generated using a validated synthetic image generator [24], and therefore, the empirical distributions are guaranteed to match the theoretical distributions.

However, to assess the quality of the datasets and to provide a tool to evaluate newly generated data, we have implemented and provided two auxiliary software scripts.

1) `test_empirical_distribution()`: Imports the generated data, plots the empirical distributions and evaluates their monotonic correlation for each combination of variable and color/shape;
2) `test_function_show()`: Plots a view of 12 images randomly selected from the entire dataset and links them to their generation parameters to validate the correct data generation procedure.

As shown in Fig. 2, using these tools, we evaluated the quality of each generated dataset by comparing the empirical distribution of each marginal against the expected theoretical one. In addition, we evaluate (and plot) the correlation between variables to verify that they correctly follow the expected multivariate distributions. Similarly, we plot the percentage of each color and shape to verify that the generated data has the desired balance of classes. Each dataset provided in this article has been evaluated according to this procedure to ensure output quality, and the corresponding plots are provided in the respective dataset folder. To ensure the quality of each generated image (i.e., to prove that each image corresponds to the correct parameter contained in the dataframe), we also manually validate 240 images randomly selected from each dataset using the `test_function_show()` script. The plots of the reference images are provided in the respective dataset folder.

RECORDS AND STORAGE

The full dataset, the synthetic data generator, and the auxiliary evaluation packages are freely available.

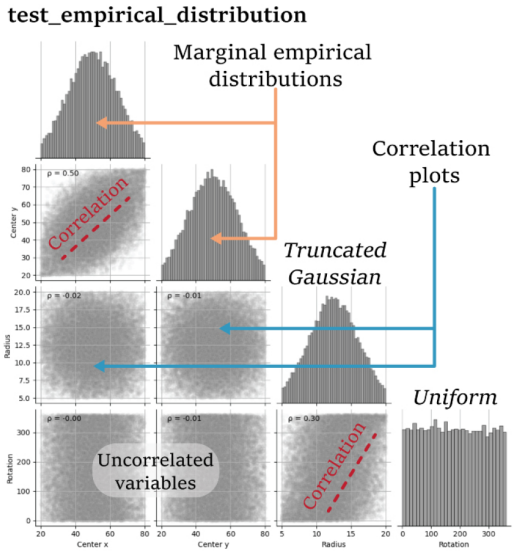

FIG. 2. **Examples of multivariate correlation analysis evaluation plots on the empirical distributions of the generated dataset.**

In particular, the dataset can be downloaded from 10.5281/zenodo.14361220 under the *Attribution-NonCommercial-NoDerivs 4.0 International* License.

According to Fig. 3, the datasets are provided in a folder containing the dataframe (`combined_dataframe.cvs`) describing for each image the unique ID and the generating parameters (i.e., the latent variables/ground truths). The synthetic data are contained in the `dataset_images` folder. To speed up file indexing and retrieval, the generated images are organized into subfolders named according to the following convention: *shape-[# sides]_color-[R-G-B]*. Each subfolder contains images stored as single PNG files, each comprising multiple tiles arranged in a 10 × 10 grid. These files are named using the format *combined_[first index]_[last index]*. The figure consists of 100 tiles arranged in ten rows and ten columns. The first tile is located at the top-left corner. Tiles are ordered column-wise: once the tiles in one column are placed from top to bottom, the next column starts from the top of the following row. To facilitate tile

loading, a Python script (*aux_data_load.py*) is included with the dataset.

The same folder contains additional data frames (`continuous_distribution_matrix.csv`, `sampler_properties_matrix.csv`, etc.) describing the properties of the dataset (e.g., number of samples, multivariate correlation matrices, and frequencies) for reproducibility. The subfolder `dataset_partial` contains the dataframes describing the images belonging to each color and shape combination.

The built-in synthetic dataset generator allows users to generate images from the dataframe that defines the properties of the dataset, eliminating the need to exchange thousands of *.png* files. In addition, the graphical interface of the dataset generator allows for easy modification of the provided dataset to better fit the research question. More details on the software implementation can be found in the associated technical report [24].

TABLE I. NADA_Dis (Disentanglement) Repository

Size	100 000/100 000/100 000	Backgr.	#000000-Black
Colors	#FFFFFF-White		
Shapes	0-Circle, 3-Triangle, 4-Square, 6-Hexagon		
Center (x)	$\mathcal{U}_{(20,80)}$ $\mathcal{N}_{(20,80)}(50,15)$ $\mathcal{N}_{(20,80)}(50,15)^*$	Center (y)	$\mathcal{U}_{(20,80)}$ $\mathcal{N}_{(20,80)}(50,15)$ $\mathcal{N}_{(20,80)}(50,15)^*$
Radius	$\mathcal{U}_{(5,20)}$ $\mathcal{N}_{(5,20)}(12.5,3.7)$ $\mathcal{N}_{(5,20)}(12.5,3.7)^{\S}$	Rotation	$\mathcal{U}_{(0,360)}$ $\mathcal{U}_{(0,360)}$ $\mathcal{U}^{\S}_{(0,360)}$
Corr.	$^*\rho = 0.5$, $^{\S}\rho = 0.3$ [Spearman]		

Note: The repository contains three datasets (100 000 elements each) of black and white images of circles, triangles, squares, and hexagons to test model disentanglement capabilities. The first dataset has uniform and independent latent variables. The second dataset contains more complex truncated Gaussian distributions, while the third adds a multivariate correlation between center positions (*) and radius/rotation (§).

[NADA_Dis] Dataset Disentanglement and Autoencoding

The main goal of modern deep architectures is to be increasingly efficient in performing classification and regression tasks, especially on high-dimensional datasets. When the training set is large, noisy, and non-homogeneous, deep architectures extract increasingly complex features and build a high-level representation to perform their regression and classification tasks. The problem with these representations is that they often lack semantic meaning, which has a negative impact on interpretability, explainability, reliability, and even performance. As Bengio [29] argued, one solution to overcome this problem is to force the networks to produce a disentangled representation of latent variables with semantic significance. Autoencoders, and especially VAE, represent a very efficient type of architecture in the disentanglement task [30], which are unsupervised or semi-supervised learning models that possess a low-dimensional latent space in which the architecture is forced to select a limited number of explanatory features capable of accurately reconstructing the output. It should be noted that the disentangling ability is highly correlated with the complexity of the intrinsic characteristics of the dataset (correlation between the variables, non-uniform distribution, etc.). However, most available datasets provide such a simplified latent representation that they do not adequately challenge novel disentangling approaches.

Therefore, we introduced *NADA_Dis* as a dataset composed of three sub-datasets of increasing latent space complexity, each consisting of 100 000 (128 × 128 pixels) unique elements (Table I). To make the disentanglement tests more effective, we selected four shapes (circles, triangles, squares, and hexagons) and three main features on which to perform controlled variations (position/center, size/radius, and rotation). All shapes are white and the background color is black. The differences between the three datasets are represented by the distributions of the selected features and the correlations of the features between them as follows (Fig. 4).

1) *NADA_Dis_1*: The shapes of the first dataset have a uniform radius distribution between 5% and 20% of the full image size ($\mathcal{U}_{(5,20)}$). Their position (Center (x), Center (y)) is evenly and randomly chosen between 20% and 80% of the total image size $\mathcal{U}_{(20,80)}$. The rotation can take place from 0° up to 360° ($\mathcal{U}_{(0,360)}$).
2) *NADA_Dis_2*: The second dataset retains the rotation, translation, and radius properties of the first but replaces the uniform distributions with truncated normal distributions with a mean of 50% and a standard deviation of 15% for the center ($\mathcal{N}_{(20,80)}(50,15)$) and a mean of 12.5 and a standard deviation of 3.7% for the radius ($\mathcal{N}_{(5,20)}(12.5,3.7)$).
3) *NADA_Dis_3*: The third dataset inherits the characteristics of the second plus a multivariate monotonic correlation of $\rho = 0.5$ between the x/y positions of the center (*) and of $\rho = 0.3$ between the radius and the rotation (§).

[NADA_OOD] Dataset Out of Distribution (OOD)

OOD detection evaluates the ability of a deep learning or machine learning model to identify inputs that do not belong to the training distributions and thus may induce unreliable predictions. Studying the behavior of a model in the presence of out-of-distributions is challenging and of great interest because it has a direct impact on the reliability and performance of the models. Recently, several methods, mainly based on the use of softmax vectors and confidence measures, have been developed for neural network image OOD detection [31]. Most OOD detection studies have been performed on public image datasets, such as Imagenet-1k [10], CIFAR-100 [32], and ResNet50 [33]. However, they have the disadvantage of not being able to systematically investigate the dynamics that most influence out-of-distribution detection, due to poor control over the

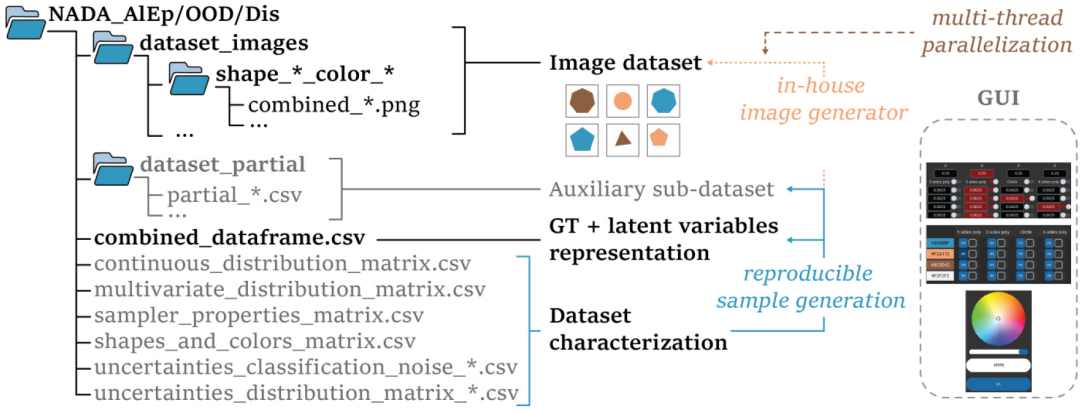

FIG. 3. Folder structure of each dataset presented. The main folder contains a subfolder (dataset_images) with the generated images. The comma-separated value file combined_dataframe.csv contains the description of each image for disentangling or training/test. The other csv files are used to fully characterize the dataset and can be loaded into the provided GUI to generate a fully reproducible variation of the dataset using an in-house image generator coupled with efficient multi-thread parallelization.

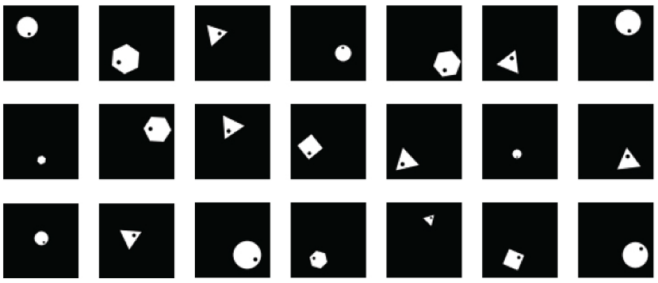

FIG. 4. NADA_Dis (disentanglement) contains images with standard (i.e., uniform) and more complex (e.g., multivariate truncated Gaussian) variable distributions to explore the disentangling ability of the model. The figures are marked with a circle to encompass the asymmetry and allow the angle of rotation to be examined.

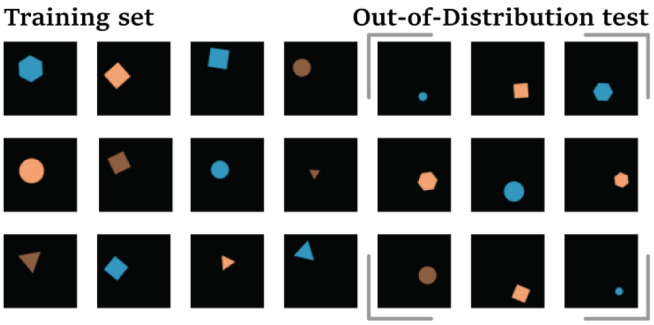

FIG. 5. Training set (100 000 elements) including larger images with the center in the upper left of the image. The NADA_OOD dataset contains six progressively different test datasets (100 000 elements each) with a progressive shift from the original distribution to the complementary one.

TABLE II. NADA_OOD (Out of Distribution) Repository

Size	100 000/6 × 100 000	Background Color	#000000-Black
Colors	#3698BF-Blue, #F2A172-Orange, #8C5D42-Brown		
Shapes	0-Circle, 3-Triangle, 4-Square, 6-Hexagon		
Probability	Blue: 0.9, 0.8, ···, 0.4 Orange: 0.05, 0.1, ···, 0.3 Brown: 0.05, 0.1, ···, 0.3		
Center (x)	$\mathcal{U}_{(20,50)}$, $\mathcal{U}_{(26,56)}$, ⋮ $\mathcal{U}_{(50,80)}$	Center (y)	$\mathcal{U}_{(20,50)}$, $\mathcal{U}_{(26,56)}$, ⋮ $\mathcal{U}_{(50,80)}$
Radius	$\mathcal{U}_{(10,20)}$, $\mathcal{U}_{(9,19)}$, ..., $\mathcal{U}_{(5,15)}$	Rotation	$\mathcal{U}_{(0,360)}$

Note: The repository contains a training dataset (100 000 images) and six test datasets (100 000 images each) with increasing differences in center position, radius, and color probability compared to the training set.

probabilistic distributions that characterize the features of the training and test sets.

For this reason, we defined *NADA_OOD* as a main training dataset of 100 000 (128 × 128 pixels) unique elements and six test datasets (100 000 unique elements each) with increasingly shifted distributions (Table II). All datasets contain shapes that are randomly selected from among: circles, triangles, squares, and hexagons. The background is black (Fig. 5).

1) *NADA_OOD_Train*: Each shape is drawn with a specific color: blue (90%), orange (5%), or brown (5%). The radius is sampled uniformly between 10% and 20% ($\mathcal{U}_{(10,20)}$) and the center of the shape is evenly distributed between 20% and 50% ($\mathcal{U}_{(20,50)}$) of the image size (setting the origin in the upper left corner).
2) *NADA_OOD_Test_*∗: The six OOD test datasets represent a progressive shift from the original color/radius/position distributions, from *NADA_OOD_Test_0*, which has the same parameters as the training set, to *NADA_OOD_Test_5*, which consists mainly of OOD images. Formally, the ith dataset has a decreasing amount of blue shapes ($p_i = 0.9 - 0.1 \cdot i$) and an increasing amount of orange and brown shapes ($p_i = (0.05 \cdot (i + 1))$). The center positions move progressively from top-left to bottom-right (center x/y position $= \mathcal{U}_{(20+6\cdot i,50+6\cdot i)}$) while decreasing their radius (radius $= \mathcal{U}_{(10-\cdot i,20-\cdot i)}$).

Models can be trained on *NADA_OOD_Train* and tested on *NADA_OOD_Test_0*. Then, on increasingly complex test

sets (monotonically increasing from *NADA_OOD_Test_1* to *NADA_OOD_Test_5*), the models can be tested for both performance degradation and capability to discriminate OOD.

[NADA_AlEp] Noisy Data, Aleatoric, and Epistemic Uncertainties

The need to extract information from noisy data and signals is a paramount concern in all scientific fields where measurement and approximation techniques are used. In particular, in the field of artificial neural networks, noisy datasets dramatically affect the accuracy, reliability, and repeatability of model performance, leading to poor prediction results, overfitting, and limited interpretability of results. Another issue that specifically affects classification tasks is label noise, as noisy labels severely degrade the generalization performance of deep neural networks. Learning noisy labels is called robust learning and is becoming an important task in modern deep learning applications [34], [35]. In the context of robust learning, Bayesian probabilistic neural networks can play a key role by quantifying the uncertainty associated with potentially incorrect labels and effectively extracting as much information as possible from datasets with reduced reliability [9]. To study the robustness of deep architectures to noise, several datasets have been created, such as natural images with synthetic noise (NISN) [36], which consists of 3000 images from Flickr corrupted with different sources of noise (Gaussian, salt and pepper, Poisson, and speckle). Other widely used datasets are the corrupted versions of MNIST (MNIST-C) [37], Cifar10 (Cifar-10C) [38], and Imagenet (Imagenet-C) [38]. Imagenet also has a perturbed version (Imagenet-P) [38]. Like Imagenet-C, Imagenet-P consists of images with 75 common visual corruptions derived from noise, blur, weather, and digital distortions. In general, quantifying the robustness and reliability of networks in the presence of noise is challenging. Moreover, quantifying the uncertainty of a model in its classification and regression tasks and distinguishing between the different types of uncertainty (aleatoric and epistemic) is an emerging research topic.

Therefore, we defined *NADA_AlEp* as a set of five synthetic datasets (100 000 unique images each, 128 × 128 pixels) with different types of noise and uncertainties (Table III). All datasets contain shapes that are randomly selected between: circles, triangles, squares, and hexagons. Colors are randomly selected from: blue, orange, and brown. The background is black. The radii can vary between 10% and 20% of the image size ($\mathcal{U}_{(10,20)}$). The center position can vary between 20% and 80% of the image size ($\mathcal{U}_{(20,80)}$) (Fig. 6).

1) *NADA_AlEp_0_Clean*: Dataset clean of noise to use as a possible training set.
2) *NADA_AlEp_1_White_Noise*: *Epistemic* white noise dataset. Each image is perturbed with an amount of white noise randomly sampled from 0% to 90% ($\mathcal{U}_{0,90}$).

TABLE III. NADA_AlEp (Aleatoric/Epistemic) Repository

Size	5 × 100 000	Background Color	#000000-Black
Colors	#3698BF-Blue, #F2A172-Orange, #8C5D42-Brown		
Shapes	0-Circle, 3-Triangle, 4-Square, 6-Hexagon		
Unc.	(0) None; (1) white noise $\mathcal{U}_{(10,90)}$; (2) deformation $\mathcal{U}_{(0,90)}$; (3) triangle $\Longleftrightarrow$ square; (4) combined (1 + 2 + 3) (random color), $p = 0.2$		
Center (x)	$\mathcal{U}_{(20,80)}$	Center (y)	$\mathcal{U}_{(20,80)}$
Radius	$\mathcal{U}_{(10,20)}$	Rotation	$\mathcal{U}_{(0,360)}$

Note: The repository contains five datasets (100 000 images each) with different amounts of aleatoric and epistemic uncertainty. (0) Standard dataset; (1) noisy dataset (white noise); (2) deformed images (original randomly stretched to circumscribed circle); (3) labeling noise (triangle to square with random color attribute); and (4) combined uncertainties.

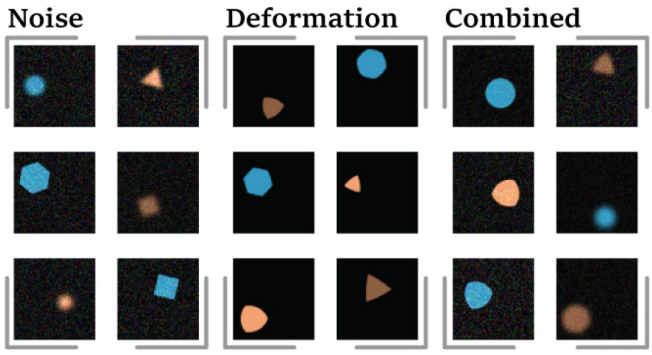

FIG. 6. NADA_AlEp (aleatoric epistemic) repository. Different types of images obtained by applying different types of uncertainties. The datasets include noisy data (obtained from white noise), deformation (from the original image to the circumscribing circle), labeling noise, and combined uncertainties to test uncertainty disentanglement.

3) *NADA_AlEp_2_Deformation*: Dataset with *Epistemic* deformation noise. Each image is deformed by a random amount, uniformly sampled between 0% and 90% ($\mathcal{U}_{(0,90)}$). Notice that 0% corresponds to the original image, while 100% is a full deformation of the circumscribing circle.
4) *NADA_AlEp_3_Label*: Dataset with label noise. Formally, 20% of triangles of a given color are misclassified as a square with a random color (among blue, orange, and brown) and vice versa (squares to triangles). Label noise introduces *aleatoric uncertainty* because it is inherent in the data and cannot be reduced.
5) *NADA_AlEp_4_Combined*: Combined dataset with all previous sources of uncertainty.

White noise and deformation are types of noise that generate epistemic uncertainty because the model, provided a large dataset and a good architecture, can reduce them. In the classification phase, the essential difference between noise that generates aleatory and epistemic uncertainty is whether or not the discriminative features that can assign an item to a particular class are preserved in the data. For example, in the case of label noise, there is an error in the ground truth, and no model improvement can reduce or resolve this

error. In the case of deformation and white noise, the data on which the model is trained are highly perturbed but retain the features that make it belong to a particular class (for example, in deformation, the number of vertices remains the fundamental feature that can distinguish a triangle from a square, even when highly perturbed). In this case, given a large dataset, the model can, by appropriately setting the hyperparameters, identify the fundamental features that can assign each element to its specific class, neglecting the superfluous or confusing ones.

INSIGHT AND NOTES

The NADA dataset and generator were developed to support the study of the behavior of probabilistic classical machine learning and deep learning architectures, to facilitate the deployment of novel methodologies, and to provide a robust benchmark for training and testing. In particular, we have identified three cutting-edge topics in the field of probabilistic supervised learning, and we have tailored three repositories (15 datasets, 100 000 images each) to study them in a systematic way. The size of the dataset is more than sufficient to train deep learning models (e.g., for shape classification, a convolutional ANN can be trained on approximately 10% of the available data, as reported in NA_DAtabase.git). Moreover, the variety introduced by the dataset's size can be leveraged to test the effects of random subsampling or the dataset-size dependence of uncertainty-aware techniques. The provided repositories allow us to investigate: the ability of networks to produce disentangled representations on high-dimensional and complex datasets (*NADA_Dis*), the ability to perform out-of-distribution detection (*NADA_OOD*), and to test the performance of a network to extract information from highly perturbed datasets affected by different types of uncertainties (*NADA_AlEp*). These are just a few examples of how NADA can be used to address research questions related to probabilistic machine learning and supervised model benchmarking. However, the integrated open-source generator (coupled with an easy-to-use graphical interface) allows each dataframe to be tailored to the specific research question or a completely new dataset to be defined from scratch given arbitrary multivariate parameter distributions.

While simple compared to most real-world applications, the datasets provide a tool to preliminary explore foundational uncertainty-aware methods in a systematic manner. Once a novel technique has been proven effective in such a simplified setting, it can be tailored to domain-specific tasks using ad hoc datasets that encompass the variability and complexity of real-world data. In future work, we will integrate conditional generative AI tools to provide more realistic textures and get closer to natural images while maintaining full control over the generation of latent variables.

SOURCE CODE AND SCRIPTS

The Python package for synthesizing data and modifying generated images can be downloaded from GitHub (NA_DAtabase.git) under the *Attribution-NonCommercial-NoDerivs 4.0 International* [24]. In addition, notebooks and implementations of several uncertainty-aware techniques can be downloaded from the GitHub repository: ShedLight_UQ.

ACKNOWLEDGMENT

The author's contributions are as follows. *Conceptualization:* GDC, CC. *Data curation and Software Implementation:* GDC, CC, FV. *Funding acquisition/Project administration:* SC, DM. The authors have declared no conflicts of interest.

REFERENCES

[1] A. S. Albahri, A. Mohammed, M. A. Fadhel, A. Alnoor, N. S. Baqer, L. Alzubaidi, O. S. Albahri, A. H. Alamoodi, J. Bai, A. Salhi, J. I. Santamaría, C. Ouyang, A. Gupta, Y. Gu, and M. Deveci, "A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion," *Inf. Fusion*, vol. 96, pp. 156–191, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257575006

[2] C. Caudai, A. Galizia, F. Geraci, L. L. Pera, V. Morea, E. Salerno, A. Via, and T. Colombo, "AI applications in functional genomics," *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 5762–5790, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:240947802

[3] G. Del Corso, R. Verzicco, and F. Viola, "Sensitivity analysis of an electrophysiology model for the left ventricle," *J. R. Soc. Interface*, vol. 17, no. 171, 2020, Art. no. 20200532.

[4] G. Del Corso et al., "Radiomics-based reliable predictions of side effects after radiotherapy for prostate cancer," in *Proc. IEEE Int. Symp. Biomed. Imaging (ISBI)*, Piscataway, NJ, USA: IEEE, 2024, pp. 1–4.

[5] G. Carloni, E. Pachetti, and S. Colantonio, "Causality-driven one-shot learning for prostate cancer grading from MRI," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 2616–2624.

[6] G. Del Corso, D. Germanese, M. A. Pascali, S. Bardelli, A. Cuttano, F. Festante, A. Guzzetta, L. Rocchitelli, and S. Colantonio, "Facial landmark identification and data preparation can significantly improve the extraction of newborns' facial features," in *Proc. IEEE 18th Int. Conf. Automat. Face Gesture Recognit. (FG)*, Piscataway, NJ, USA: IEEE, 2024, pp. 1–7.

[7] G. Del Corso et al., "Adaptive machine learning approach for importance evaluation of multimodal breast cancer radiomic features," *J. Imaging Inform. Med.*, vol. 37, pp. 1–10, 2024.

[8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, pp. 139–144, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:1033682

[9] L. V. Jospin, W. L. Buntine, F. Boussaid, H. Laga, and Bennamoun, "Hands-on Bayesian neural networks—A tutorial for deep learning users," *IEEE Comput. Intell. Mag.*, vol. 17, no. 2, pp. 29–48, May 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:220514248

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Piscataway, NJ, USA: IEEE, 2009, pp. 248–255.

[11] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361. [Online]. Available: https://api.semanticscholar.org/CorpusID:6724907

[12] P. Roy, S. Ghosh, S. Bhattacharya, and U. Pal, "Effects of degradations on deep neural network architectures," 2018, *arXiv:1807.10108*.

[13] L. Deng, "The MNIST database of handwritten digit images for machine learning research," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 141–142, Nov. 2012.

[14] G. Cohen, S. Afshar, J. C. Tapson, and A. van Schaik, "EMNIST: An extension of MNIST to handwritten letters," 2017, *arXiv:1702.05373*.

[15] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," 2017, *arXiv:1708.07747*.

[16] L. Matthey, I. Higgins, D. Hassabis, and A. Lerchner, "dSprites: Disentanglement testing sprites dataset," 2017. [Online]. Available: https://github.com/deepmind/dsprites-dataset/

[17] A. E. Korchi and Y. Ghanou, "2D geometric shapes dataset—For machine learning and pattern recognition," *Data Brief*, vol. 32, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221007484

[18] N. Michlo, R. Klein, and S. D. James, "Overlooked implications of the reconstruction loss for VAE disentanglement," in *Proc. Int. Joint Conf. Artif. Intell.*, 2022, pp. 4073–4081. [Online]. Available: https://api.semanticscholar.org/CorpusID:250089449

[19] N. Singh, "600 synthetic shapes: Circles, squares, triangles," 2023. [Online]. Available: https://www.kaggle.com/datasets/singhnavjot2062001/geometric-shapes-circle-square-triangle

[20] J. Hou, C. Yan, R. Li, Q.-L. Huang, X. Fan, and F. Lin, "Lung nodule segmentation algorithm with SMR-UNET," *IEEE Access*, vol. 11, pp. 34319–34331, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258002943

[21] T. Mimori, K. Sasada, H. Matsui, and I. Sato, "Diagnostic uncertainty calibration: Towards reliable machine predictions in medical domain," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:220347273

[22] K. Sasada, N. Yamamoto, H. Masuda, Y. Tanaka, A. Ishihara, Y. Takamatsu, Y. Yatomi, W. Katsuda, I. Sato, and H. Matsui, "Inter-observer variance and the need for standardization in the morphological classification of myelodysplastic syndrome." *Leukemia Res.*, vol. 69, pp. 54–59, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:4891350

[23] J. A. van der Putten, F. van der Sommen, J. de Groof, M. R. Struyvenberg, S. Zinger, W. L. Curvers, E. J. Schoon, J. J. Bergman, and P. H. N. D. with, "Modeling clinical assessor intervariability using deep hypersphere encoder–decoder networks," *Neural Comput. Appl.*, vol. 32, pp. 10705–10717, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:208211214

[24] F. Volpini, C. Caudai, G. Del Corso, and S. Colantonio, "Na database: Generator of probabilistic synthetic geometrical shape dataset," *ISTI-CNR: Techn. Rep.*, 2024. [Online]. Available: https://iris.cnr.it/bitstream/20.500.14243/505804/1/Report_tecnico_NaDA.pdf

[25] G. D. Corso, S. Colantonio, and C. Caudai, "Shedding light on uncertainties in machine learning: Formal derivation and optimal model selection," *J. Frankl. Inst.*, vol. 362, 2025, Art. no. 107548. [Online]. Available: https://api.semanticscholar.org/CorpusID:275792987

[26] A. Sklar, "Random variables, joint distribution functions, and copulas," *Kybernetika*, vol. 9, no. 6, pp. 449–460, 1973.

[27] K. Conrad, "Probability distributions and maximum entropy," *Entropy*, vol. 6, no. 452, p. 10, 2004.

[28] M. Baudin, A. Dutfoy, B. Iooss, and A.-L. Popelin, *OpenTURNS: An Industrial Software for Uncertainty Quantification in Simulation.* Cham, Germany: Springer International Publishing, 2016, pp. 1–38. [Online]. Available: https://doi.org/10.1007/978-3-319-11259-6_64-1

[29] Y. Bengio, "Deep learning of representations: Looking forward," in *Proc. Int. Conf. Statist. Language Speech Process.*, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:1044293

[30] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[31] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:3526391

[32] A. Krizhevsky, "Learning multiple layers of features from tiny images," 2009. [Online]. Available: https://api.semanticscholar.org/CorpusID:18268744

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 770–778. [Online]. Available: https://api.semanticscholar.org/CorpusID:206594692

[34] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 34, no. 11, pp. 8135–8153, Nov. 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:220546541

[35] L. Jiang, D. Huang, M. Liu, and W. Yang, "Beyond synthetic noise: Deep learning on controlled noisy labels," in *Proc. Int. Conf. Mach. Learn.*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:220265487

[36] T. Pathak, "Natural images with synthetic noise," 2021. [Online]. Available: https://www.kaggle.com/dsv/1958494

[37] N. Mu and J. Gilmer, "Mnist-c: A robustness benchmark for computer vision," 2019, *arXiv:1906.02337*.

[38] D. Hendrycks and T. G. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," 2019, *arXiv:1903.12261*.