Inductive Conformal Out-of-distribution Detection based on Adversarial Autoencoders

Feiyang Cai, Ali I. Ozdagli, Nicholas Potteiger and Xenofon Koutsoukos
Institute for Software Integrated Systems, *Vanderbilt University*
Email: {feiyang.cai, ali.i.azdagli, nicholas.potteiger, xenofon.koutsoukos}@vanderbilt.edu

2021 IEEE International Conference on Omni-Layer Intelligent Systems (COINS) | 978-1-6654-3156-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/COINS51742.2021.9524167

***Abstract*—Machine learning components are used extensively to cope with various complex tasks in highly-uncertain environments. However, Out-Of-Distribution (OOD) data may lead to predictions with large errors and degrade performance considerably. This paper first introduces different types of OOD data and then presents an approach for OOD detection for classification problems efficiently. Our approach utilizes an Adversarial Autoencoder (AAE) for representing the training distribution and Inductive Conformal Anomaly Detection (ICAD) for online detecting OOD high-dimensional data. Experimental results using several datasets demonstrate that the approach can detect various types of OOD data with a small number of false alarms. Moreover, the execution time is very short, allowing for online detection.**

***Keywords*—Out-of-distribution detection, Machine learning components for classification, Adversarial autoencoder, Inductive conformal anomaly detection.**

I. Introduction

Over the past decade, machine learning components, such as Deep Neural Networks (DNNs), have made remarkable achievements, resulting in state-of-the-art performance in various tasks, especially in image classification systems [1], [2]. Nevertheless, there are still several challenges restricting the deployment of machine learning components to safety-critical real-world systems. Machine learning models are built upon an underlying assumption that the training and testing data are sampled from the same distribution. In a real-world system, however, even if a machine learning component is well-trained over an extensive training dataset, Out-Of-Distribution (OOD) data are still inevitable during testing, and they may cause the model to make erroneous predictions and degrade the performance considerably. Hence, detection of OOD data is significant for the safety of machine learning components. When OOD data are fed into the predictive model, the detector can raise alarms for human intervention or redesign of the model.

Although there are many studies on OOD detection in machine learning components, especially for classification, the manifestations of OOD data are still unexplored. OOD detection methods in the literature [3], [4] attempt to determine whether an input example is from the same distribution as the training dataset, which only detects the change in the distribution of the input variable. Another research direction is novelty detection for unknown classes [5], [6]. The novelties from unknown classes can be regarded as another type of OOD data, where the change in the distribution of the output variable is also observed. A related research topic to OOD data is *dataset shift*, which occurs when the joint distribution of the input and output variables differ between the training and testing phases [7]. This paper analyzes the causes of out-of-distribution data and categorizes them into three types: OOD data caused by covariate shift, label shift, and concept shift. It should be noted that, we borrow the idea and terminologies from the categories of dataset shifts. However, there is still a specific difference between dataset shift and OOD data: dataset shift focuses on two distributions – the distributions of the training dataset and testing dataset; in contrast, the OOD data focuses on the distribution of the training dataset and a single test example.

In order to efficiently detect different types of OOD data in machine learning components, we propose the inductive conformal out-of-distribution detection, which is based on the Inductive Conformal Anomaly Detection (ICAD) framework [8]. The core of the ICAD method is the definition of a *nonconformity measure*, which is a function measuring the dissimilarity between a test example and the training dataset. Our approach utilizes a variant of an Adversarial Autoencoder (AAE) [9] to define the nonconformity measure, which can disentangle the label information from the latent representation by estimating a class variable in addition to the latent representation. By using such an architecture, the joint distribution of the input and output variables on the training dataset can be represented. Therefore, both the input and output of the machine learning component can be taken into consideration for OOD detection.

Moreover, the detection method using a single example may result in a large number of false alarms. The robustness of the detector can be improved by incorporating multiple examples into the detection algorithm [10]. Our method follows this idea and employs an AAE to generate multiple examples for robust detection. Although multiple examples are considered, our approach focuses on comparing a single test example with the training distribution nonetheless. We also design two different nonconformity measures, quantifying the degree to which the test example is not sampled from the same distribution as the training dataset. We conduct extensive experiments on several datasets to evaluate our approach. The results show that our approach can efficiently detect different types of OOD data and can be used for online detection.

The rest of this paper is organized as follows: Section II

formulates the problem of OOD detection and discusses different types of OOD data in machine learning components for classification. Section III describes our proposed approach – inductive conformal out-of-distribution detection. Section IV utilizes multiple datasets to demonstrate our detection method. Section V discusses the related work, and Section VI provides the concluding remarks.

II. Out-of-distribution Detection in Machine Learning Components for Classification

In this section, we first formulate the OOD detection problem in machine learning components for classification. After that, we categorize the OOD data according to their cause.

A. Problem Formulation

Consider a machine learning component f for a classification problem, which is well-trained using a set of labeled samples $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{l}$, where each example (x_i, y_i) consists of the input $x_i \in \mathcal{X}$ and corresponding label $y_i \in \mathcal{Y}$, and it is sampled from a joint distribution $P_{\text{train}}(x, y)$. During the system operation, a test example x_{l+1} is consumed by the component to estimate a predictive class y'_{l+1}. The implicit assumption for the effectiveness of machine learning techniques is that the test example pair (x_{l+1}, y'_{l+1}) is sampled from the same joint distribution of the training dataset $P_{\text{train}(x,y)}$. However, the training dataset $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{l}$ is necessarily incomplete, and therefore, OOD data are commonly present. The machine learning component f may become ineffective due to the OOD data and make predictions with large errors. In this case, it is desirable to raise alarms or retrain the model, and therefore, detection of OOD data is of importance for the safety of the machine learning component.

The detection must be performed efficiently and preferably online, which means that the execution time should be comparable to the execution time of the machine learning component. It is very challenging because machine learning components are increasingly used for tasks with high-dimensional data.

B. Types of Out-of-distribution Data

1) Out-of-distribution data caused by covariate shift: Covariate shift is one of the basic and most common dataset shifts observed in real life [11]. Covariate shift usually occurs when $P_{\text{test}}(x)$ changes over time after training, while the conditional probability $P(y|x)$ remains the same. Extending this definition to OOD data, if the input variable x of a test example is not sampled from the same distribution of training dataset $P_{\text{train}}(x)$, while the underlying relationship between input and output $P(y|x)$ remains unchanged, it is assumed that the OOD data are caused by covariate shift.

2) Out-of-distribution data caused by label shift: Label shift, also known as target shift, assumes that the marginal distribution of y, $P(y)$ changes, but everything else remains the same [7]. For label shift, the training and testing distributions of output variables may change in time such that $P_{\text{train}}(y) \neq P_{\text{test}}(y)$. However, the conditional probability of x given y stays same, i.e. $P_{\text{train}}(x|y) = P_{\text{test}}(x|y)$. Subsequently, when the label variable y is out of distribution of the training dataset $P_{\text{train}}(y)$, but the underlying relationship $P(x|y)$ remains the same, it is assumed that the OOD data are caused by label shift.

3) Out-of-distribution data caused by concept shift: A concept shift is a form of contextual shift where the relationship between input and output variables changes [12]. Here, we assume the data generation mechanism, $P(x|y)$ changes while class definitions remain same. For a test example, if its output variable y is in the same distribution of the training dataset $P\text{train}(y)$, but the conditional probability $P(x|y)$ changes, we define such an example as OOD data caused by concept shift.

III. Adversarial Autoencoder and Out-of-distribution Detection

In this section, we introduce inductive conformal out-of-distribution detection, which is based on a variant of an Adversarial Autoencoder (AAE) and the Inductive Conformal Anomaly Detection (ICAD) framework.

A. Adversarial Autoencoder

An AAE is a generative model which is trained in an adversarial manner to force the aggregated posterior of the latent coding space of the autoencoder to match an arbitrary known distribution [9]. Specifically, assuming x is the input and z is the low-dimensional latent representation, a basic AAE model consists of an encoder (generator in adversarial network) $G(x)$ trying to encode the input into the low-dimensional latent representation, a decoder $De(z)$ trying to reconstruct the original input data from the encodings, and a discriminator $D(z)$ trying to identify the hidden samples z generated by the generator or sampled from the true prior. The whole architecture is trained jointly in two phases: the *reconstruction* phase and *regularization* phase. In the reconstruction phase, the encoder $G(x)$ and decoder $De(z)$ are updated to minimize the reconstruction error. In the regularization phase, the discriminator $D(z)$ is trained to distinguish the true samples (sampled from the prior distribution $p(z)$) from the generated samples (sampled from the posterior distribution $q(z|x)$), while the generator $G(x)$ is trained to deceive the discriminator $D(z)$ by outputting samples that closely resemble data sampled from the prior distribution $p(z)$.

In order to disentangle the label information from the latent representation, a class variable y can be predicted by the encoder $G(x)$ in addition to the latent variable z, and the one-hot vector of the predicted class is provided to the decoder $De(y, z)$ to generate class-conditioned output (Fig. 1). This architecture can be regarded as a supervised variant of a semi-supervised AAE introduced in [9]. A *supervised classification* phase is performed after the reconstruction and regularization phases, whose objective is to minimize the cross-entropy cost between the target distribution $p(y)$ and the approximation of the target distribution $q(y|x)$.

B. Inductive Conformal Out-of-distribution Detection

The task of anomaly detection is to determine whether the test example conforms to the normal data. Inductive Conformal

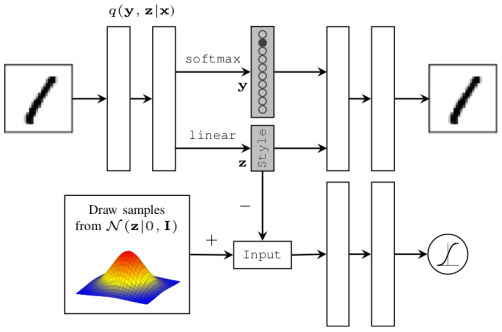

Fig. 1. A variant of the adversarial autoencoder model.

Anomaly Detection (ICAD) is an anomaly detection framework with the property of *well-calibrated false alarms* [8]. The ICAD method is based on a definition of a *nonconformity measure*, which is a function measuring the dissimilarity between a test example and the training dataset. Recently, in order to enable online detection for high-dimensional data, learning models are trained to represent the distribution of the input variable of the training dataset and are utilized for the computation of a nonconformity measure [10]. However, nonconformity measures considering only the input variable may not be sufficient for the detection of some specific types of OOD data because the output variable can also lead the input and output pair out of the joint distribution of the training dataset. The variant of AAE can encode a label variable in addition to a latent variable, and consequently, both input and output can be represented in such a model. In the following, we first introduce two different nonconformity measures based on AAE. Subsequently, we describe the inductive conformal OOD detection method based on these nonconformity measures.

1) Nonconformity measures: For a test example x, the encoder portion of the AAE represents x and its predictive label y' in a latent space, and the decoder portion generates a new example $\hat{x}$ by sampling from the encodings. If x and its predictive label y' are from the same joint distribution of the training dataset, the example x should be reconstructed with a relatively small reconstruction error. Therefore, the reconstruction error between the input x and generated output $\hat{x}$ can be used as the reconstruction-based nonconformity measure A_{rc} defined as

$$A_{\text{rc}} = ||x - \hat{x}||^2. \tag{1}$$

The reconstruction-based nonconformity measure treats all features of the input equally. However, a relatively small part of the features in the input may have a significant effect on the final prediction. Therefore, it is not reasonable to treat all input features equally when they contribute to the output of the predictive model differently. A novel nonconformity measure based on *saliency maps* is introduced to compensate for such a defect. A saliency map algorithm aims to quantify the contributions of the input features to the predictive result of a machine learning model [13]. Specifically, we utilize the gradient-based saliency map algorithm [13]. It generates the saliency map by computing the derivative of the SoftMax score S_y of class y with respect to the input x at a given point x_0

$$w = \frac{\partial S_y}{\partial x}|_{x_0}.$$

The derivative w reflects the influence of input features on the final prediction and is used to define the saliency-based nonconformity measure by weighting the reconstruction error as

$$A_{\text{saliency}} = ||w \cdot (x - \hat{x})||^2. \tag{2}$$

The saliency map w is computed using the portion used for the classification task in the encoder of the AAE, and the reconstructed output $\hat{x}$ is also generated by the AAE.

2) Detection method: Given a test input x_{l+1} and its predictive label y'_{l+1}, the OOD detection method aims to determine whether the test input-prediction pair (x_{l+1}, y'_{l+1}) is sampled from the same joint distribution of the training dataset $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{l}$. The proposed method is based on the framework of ICAD, and therefore, the detection algorithm is divided into offline and online phases. During the offline phase, the training dataset $\mathcal{D}_{\text{train}}$ is split into two sets: a proper training set $\mathcal{D}_{\text{proper}} = \{(x_i, y_i)\}_{i=1}^{m}$ and a calibration set $\mathcal{D}_{\text{calibration}} = \{(x_i, y_i)\}_{i=m+1}^{l}$. An AAE F is trained over a proper training set $\mathcal{D}_{\text{proper}}$ for the computation of nonconformity measures. Let A be either nonconformity measure function defined before. After that, for each data $x_j, j = m+1, \ldots, l$ in the calibration set, a new example $\hat{x}_j$ is generated using the trained AAE F, and its corresponding nonconformity score α_j^{Γ} is computed according to the nonconformity measure A. In order to reduce the time complexity of the p-value computation during the online phase, nonconformity scores of the calibration data are sorted and stored as $\{\alpha_j\}_{j=m+1}^{l}$.

At the online detection stage, given the test example x_{l+1}, in order to improve the robustness of the detection, N examples $\{\hat{x}_{l+1,k}\}_{k=1}^{N}$ are generated from the AAE. For each generated example $\hat{x}_{l+1,k}$, its nonconformity score $\alpha_{l+1,k}$ can be computed using the same nonconformity measure A as the calibration set. Subsequently, two different techniques can be applied to aggregate these N nonconformity scores for detection.

One option is to compute the expected nonconformity score $\bar{\alpha}_{l+1}$ of N nonconformity scores, and the p-value p_{l+1} can be computed as the ratio of calibration nonconformity scores that are at least as large as $\bar{\alpha}_{l+1}$:

$$p_{l+1} = \frac{|\{i = m+1, \ldots, l\} \,|\, \alpha_i \geq \bar{\alpha}_{l+1}|}{l - m}. \tag{3}$$

A smaller p-value reflects an unusual test example with respect to the training examples. If the p-value p_{l+1} is smaller than a threshold ϵ, this test example will be classified as an OOD instance.

Additionally, we can use a martingale test [14], [10] for N nonconformity scores to detect OOD data. For each nonconformity score of a generated example $\alpha_{l+1,k}$, the corresponding

p-value $p_{l+1,k}$ is calculated using Eq. (3). Then, a simple mixture martingale [14] is applied, which is defined as

$$M_{l+1} = \int_0^1 \prod_{k=1}^{N} \epsilon p_{l+1,k}^{\epsilon-1} d\epsilon. \tag{4}$$

Such martingale value will grow only if there are many small p-values in $\{p_{l+1,k}\}_{k=1}^{N}$, and the detector will raise an alarm when the martingale value M_{l+1} is greater than a predefined threshold τ. The martingale test is expected to have a better performance than the expected p-value of nonconformity scores since it can enlarge the nonconformity gap between in-distribution and OOD instances.

The whole procedure of the detection algorithm is summarized in Algorithm 1.

Algorithm 1 Inductive Conformal Out-of-distribution Detection using Adversarial Autoencoders

Require: Input training set $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{l}$; number of calibration examples m; number of examples N generated by the adversarial autoencoder; test example x_{l+1}; threshold ϵ of p-value of expected nonconformity score, or threshold τ of martingale value

Offline:

Split the training set $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{l}$ into the proper training set $\mathcal{D}_{\text{proper}} = \{(x_i, y_i)\}_{i=1}^{l}$ and calibration set $\mathcal{D}_{\text{proper}} = \{(x_i, y_i)\}_{i=m+1}^{l}$
Train an adversarial autoencoder F using the proper training set $\mathcal{D}_{\text{proper}}$
for $j = m+1$ to l **do**
    Generate $\hat{x}_j$ using the trained adversarial autoencoder
    $\alpha_j^{\Gamma} = A(x_j, \hat{x}_j)$
end for
$\{\alpha_{m+1}, \ldots, \alpha_l\}$ = sort($\{\alpha_{m+1}^{\Gamma}, \ldots, \alpha_l^{\Gamma}\}$)

Online (p-value of expected nonconformity score):

for $k = 1$ to N **do**
    Generate $\hat{x}_{l+1,k}$ using the trained adversarial autoencoder
    $\alpha_{l+1,k} = A(x_{l+1}, \hat{x}_{l+1,k})$
end for
$\bar{\alpha}_{l+1} = \frac{1}{N}\sum_{k=1}^{N} \alpha_{l+1,k}$
$p_{l+1} = \frac{|\{i=m+1,\ldots,l\} \mid \alpha_i \geq \bar{\alpha}_{l+1}|}{l-m}$
$Anom_{l+1} \leftarrow p_{l+1} > \epsilon$

Online (martingale test):

for $k = 1$ to N **do**
    Generate $\hat{x}_{l+1,k}$ using the trained adversarial autoencoder
    $\alpha_{l+1,k} = A(x_{l+1}, \hat{x}_{l+1,k})$
    $p_{l+1,k} = \frac{|\{i=m+1,\ldots,l\} \mid \alpha_i \geq \alpha_{l+1,k}|}{l-m}$
end for
$M_{l+1} = \int_0^1 \prod_{k=1}^{N} \epsilon p_{l+1,k}^{\epsilon-1} d\epsilon$
$Anom_{l+1} \leftarrow M_{l+1} > \tau$

IV. Evaluation

To demonstrate the effectiveness of the proposed approach, we conduct extensive experiments for the detection of different types of OOD data using several datasets. In this section, we describe the implementation details first. Then, we describe the experimental setup and present evaluation results for three different types of OOD. Finally, we measure and report the execution time of the proposed method.

A. Experiment Implementation

1) Neural network architecture: The AAE is trained to perform both classification and detection tasks. For different types of inputs, we use different architectures of the AAE. For the image input (Experiment 1-2), in order to allow for online detection, we implement the AAE with a relatively shallow convolutional network: the encoder contains three convolutional layers and one fully connected layer. The decoder has symmetric one fully connected layer and three deconvolutional layers. Furthermore, three fully connected layers form the discriminator. For the non-image input (Experiment 3), the AAE is implemented with three fully connected layers. The decoder has a symmetric architecture, and the discriminator contains three fully-connected layers.

2) Evaluation metrics: The Receiver Operating Characteristic (ROC) curve plots the true positive rate against the false positive rate by varying the detection threshold. The Area Under ROC (AUROC) curve is a threshold-free metric and is considered as the evaluation metric for OOD detection. The worst value of AUROC is 0.5 yielded by an uninformative classifier with a random guess. The best value of AUROC is 1.0, implying that the nonconformity scores for all the OOD data are greater than the score for any in-distribution data.

B. Out-of-distribution Data Caused by Covariate Shift

Experiment 1: The OOD data caused by covariate shift is present when the input variable x is sampled from a different distribution of training dataset, but the underlying relationship between the input and output $P(y|x)$ remains unchanged. MNIST [15], colorful MNIST [16], and SVHN [17], are the image classification datasets with the same labels of ten digits, while the inputs are from different distributions: for MNIST, the inputs are black and white images with handwritten digits; for colorful MNIST, the inputs are the MNIST images synthesized with colorful backgrounds; for SVHN, the inputs are the digit images from the street view house numbers. In our experiment, we train the AAE with the MNIST dataset and test it with colorful MNIST (Experiment 1-1) and SVHN (Experiment 1-2). It can be regarded as the OOD data caused by covariate shift since the input variable is sampled from a different distribution of training dataset, but the conditional probability of y given x stays the same.

Results of Experiment 1: We report the accuracy of the classification task and the AUROC of the detection task in Table I using different nonconformity measures and different techniques applied to nonconformity scores ("Ave" is for the technique using expected p-value of N nonconformity scores; "Mart" is for the technique using martingale test). As it can be seen from the table, the accuracy of the classification for the in-distribution data is not degraded. Further, the AUROC in

Experiment 1-1 and 1-2 are almost close to 1.0. Therefore, the proposed method can be used to detect the out-of-distribution data caused by covariate shift. Besides, it should be noted that for this experiment, no baselines can be compared in the literature.

TABLE I
AUROC FOR INDUCTIVE OUT-OF-DISTRIBUTION DETECTION.

	Accuracy	A_{rc}(Ave/Mart)	A_{saliency}(Ave/Mart)
Experiment 1-1	99.3%	0.998/0.999	0.999/0.999
Experiment 1-2		1.000/1.000	1.000/1.000
Experiment 2-1	99.5%	0.943/0.932	0940/0.931
Experiment 2-2	97.2%	0.840/0.847	0.829/0.821
Experiment 2-3	90.1%	0.692/0.683	0.683/0.690
Experiment 3	96.3%	0.682/0.693	0.674/0.681

C. Out-of-distribution Data Caused by Label Shift

Experiment 2: Novelty detection for unknown classes is a representative example of the OOD data caused by label shift, where the label variable y is not sampled from the same distribution of the training dataset, but the output-conditional distribution $P(x|y)$ remains the same. In our experiment, following the experimental settings in [6], we randomly sample 6 classes in MNIST(Experiment2-1) [15], SVHN(Experiment 2-2) [17], and CIFAR10(Experiment 2-3) [18] as known classes, and rest 4 classes are unknowns. The training dataset only contains the 6 known classes, but the testing dataset contains all 10 classes.

Results of experiment 2: In this case, we report the evaluation results in Table I. The results demonstrate the effectiveness of our approach for detecting the OOD data caused by label shift. Although the AUROC of our approach is smaller than the other state-of-the-art methods [6], [19], our approach uses a more shallow neural network allowing for online detection.

D. Out-of-distribution Data Caused by Concept Shift

Experiment 3: The OOD data caused by concept shift is present when the output variable y is sampled from the training dataset, but the conditional probability of $P(x|y)$ changes. A gear dataset is used in the experiment to evaluate our approach for detecting such OOD data. Gearbox fault detection dataset [20] focuses on classifying the type of damage that may occur on a generic gearbox. The state of the gearbox is measured using accelerometers attached at various locations. The gearbox can operate at five different constant shaft speeds under two different loading conditions (low- and high-load). For each shaft speed and loading conditions, six fault types are simulated (normal, chipped gear tooth, broken gear tooth, bent shaft, imbalanced shaft, broken gear tooth with bent shaft). For each case which is the combination of fault type, shaft speed, and load condition, about 4 seconds of data are collected at a sampling rate of 66.67 kHz twice. To make the dataset suitable for this research, preprocessing is performed: In this paper, only the output shaft vibration data is considered. The dataset is divided into two main subsets; those are low-load and high-load. For each subset, regardless of shaft speed, the dataset is aggregated with respect to the type of fault. All available data is converted into the frequency domain using Short Time Fourier Transform. Our experiment uses the subset from the low-load as the training dataset and both subsets from low- and high-load as the testing dataset. The distribution over output $P(y)$ stays the same, but the conditional probability $P(x|y)$ changes. Thus, it can be regarded as the OOD data caused by concept shift.

Results of experiment 3: Evaluation results corresponding to the experiment are shown in Table I. Although there is no baseline in the literature used to compare with, as it can be seen from this table, the approach can detect OOD data caused by concept shift using different nonconformity measures without loss of classification accuracy.

E. Execution Time

In order to characterize the efficiency of the approach, we measure and report execution times for Experiment 1-1 using two different nonconformity measures and martingale test using a box plot in Fig. 2. The execution times are measured on a 6-core Ryzen 5 desktop with a single GTX 1080Ti GPU.

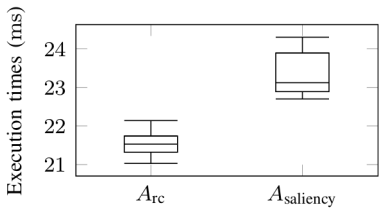

Fig. 2. Execution times of proposed method.

From the results, the execution times of the method using nonconformity measure A_{saliency} are slightly longer than the method using A_{rc} due to the extra execution time for computing the saliency maps. The execution times for all two nonconformity measures are very short, which is comparable to the inference time of typical machine learning component [21].. Therefore, our approach is applicable for online detection. Moreover, the number of the examples generated from AAE N is fixed at 10 in experiments. As the number of N increases, the execution time will also increase since the AAE model needs to be inferred N times to generate N examples.

V. RELATED WORK

In the last decade, a significant amount of work in the literature focuses on detecting OOD examples. A baseline method is proposed in [3], which utilizes the SoftMax score to classify the in-distribution and OOD instances. A method of learning confidence estimates for neural networks is proposed in [22], which adds an additional branch to yield a confidence logit. The OOD detection can be performed by evaluating the learned confidence estimates.

Data from unknown classes can be considered as a type of out-of-distribution data. The problem of detecting the novelty for unknown classes is usually considered together with open-set recognition. Open-set recognition performs two tasks:

novelty detection for unknown classes and classification for known classes. A Class Conditioned Autoencoder (C2AE) is trained for open-set recognition in [6]. Conditional Gaussian Distribution Learning (CGDL) method is presented in [23], which applies a probabilistic ladder network trying to learn conditional Gaussian distributions by forcing different latent features to approximate different Gaussian models. The reconstruction error and the probability of the test sample located in the latent space are combined to detect the unknown class.

When the distribution of the test dataset is given instead of only a single test example, the problem of OOD detection evolves into dataset shift detection. The state-of-the-art approaches for label shift detection known as Black Box Shift Estimation (BBSE) and Maximum Likelihood Label Shift (MLLS) are proposed in [24] and [25], respectively. As for covariate shift, an Exponentially Weighted Moving Average chart (EWMA) model is used in [26] to detect covariate shifts in non-stationary environments. ADaptive WINdowing (ADWIN), an adaptive sliding window algorithm for detecting concept shift, or concept drift, is raised in [27]. In [28], the conformal prediction and exchangeability martingales are adapted for testing concept shift online.

VI. Conclusions

In this paper, we formalize the problem of detecting OOD data in machine learning components for classification and categorize the OOD data according to their causes. Then, we present an approach based on inductive conformal anomaly detection. An adversarial autoencoder model is adopted to characterize the joint distribution of the training dataset, allowing online detection for high-dimensional data. Experiments using several datasets demonstrate the effectiveness of the approach for the detection of different types of OOD data. Moreover, the execution time is very small, and consequently, the approach can be used for online out-of-distribution detection. To extend this work, we plan to compare our approach with state-of-the-art methods for OOD detection to demonstrate the benefit of taking the output into consideration. Evaluation with real-world image datasets is also a part of future work.

Acknowledgement

The material presented in this paper is based upon work supported by the National Science Foundation (NSF) under grant numbers CNS 1739328 and the Defense Advanced Research Projects Agency (DARPA) through contract number FA8750-18-C-0089. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, or NSF.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 26th Annual Conference on Neural Information Processing Systems*, 2012.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

[3] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proceedings of the 5th International Conference on Learning Representations*, 2017.

[4] D. Hendrycks, M. Mazeika, and T. G. Dietterich, "Deep anomaly detection with outlier exposure," in *Proceedings of the 7th International Conference on Learning Representations*, 2019.

[5] A. Bendale and T. E. Boult, "Towards open set deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[6] P. Oza and V. M. Patel, "C2AE: class conditioned auto-encoder for open-set recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[7] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset shift in machine learning*. The MIT Press, 2009.

[8] R. Laxhammar and G. Falkman, "Inductive conformal anomaly detection for sequential detection of anomalous sub-trajectories," *Annals of Mathematics and Artificial Intelligence*, vol. 74, no. 1-2, pp. 67–94, 2015.

[9] A. Makhzani, J. Shlens, N. Jaitly, and I. Goodfellow, "Adversarial autoencoders," in *Workshop Track Proceedings of the 4th International Conference on Learning Representations*, 2016.

[10] F. Cai and X. D. Koutsoukos, "Real-time out-of-distribution detection in learning-enabled cyber-physical systems," in *Proceedings of the 11th ACM/IEEE International Conference on Cyber-Physical Systems, ICCPS*. IEEE, 2020.

[11] M. Sugiyama, M. Krauledat, and K.-R. MÃžller, "Covariate shift adaptation by importance weighted cross validation," *Journal of Machine Learning Research*, vol. 8, no. May, pp. 985–1005, 2007.

[12] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern recognition*, vol. 45, no. 1, pp. 521–530, 2012.

[13] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Workshop Track Proceedings of the 2nd International Conference on Learning Representations*, 2014.

[14] V. Fedorova, A. J. Gammerman, I. Nouretdinov, and V. Vovk, "Plug-in martingales for testing exchangeability on-line," in *Proceedings of the 29th International Conference on Machine Learning*, 2012.

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[16] "Getting started with gans part 2: Colorful mnist," https://www.wouterbulten.nl/blog/tech/getting-started-with-gans-2-colorful-mnist/.

[17] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *Workshop Track Proceedings of the 25th Annual Conference on Neural Information Processing Systems*, 2011.

[18] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[19] G. Chen, P. Peng, X. Wang, and Y. Tian, "Adversarial reciprocal points learning for open set recognition," *arXiv preprint*.

[20] "Phm data challenge," http://www.phmsociety.org/competition/09, 2009, accessed: 2009-09-28.

[21] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint*, 2018.

[22] T. DeVries and G. W. Taylor, "Learning confidence for out-of-distribution detection in neural networks," *arXiv preprint*, 2018.

[23] X. Sun, Z. Yang, C. Zhang, K. V. Ling, and G. Peng, "Conditional gaussian distribution learning for open set recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[24] Z. C. Lipton, Y.-X. Wang, and A. Smola, "Detecting and correcting for label shift with black box predictors," *arXiv preprint*, 2018.

[25] S. Garg, Y. Wu, S. Balakrishnan, and Z. C. Lipton, "A unified view of label shift estimation," *arXiv preprint arXiv:2003.07554*, 2020.

[26] H. Raza, G. Prasad, and Y. Li, "Ewma model based shift-detection methods for detecting covariate shifts in non-stationary environments," *Pattern Recognition*, vol. 48, no. 3, pp. 659–669, 2015.

[27] A. Bifet and R. Gavalda, "Learning from time-changing data with adaptive windowing," in *Proceedings of the 2007 SIAM international conference on data mining*. SIAM, 2007, pp. 443–448.

[28] V. Vovk, "Testing for concept shift online," *arXiv preprint*, 2020.