PAC-Based Formal Verification for Out-of-Distribution Data Detection

Mohit Prashant
School of Computer Science and Engineering
NTU, Singapore
mohit010@e.ntu.edu.sg

Arvind Easwaran
School of Computer Science and Engineering
NTU, Singapore
arvinde@ntu.edu.sg

Abstract—Cyber-physical systems (CPS) like autonomous vehicles, that utilize learning components, are often sensitive to noise and out-of-distribution (OOD) instances encountered during runtime. As such, safety critical tasks depend upon OOD detection subsystems in order to restore the CPS to a known state or interrupt execution to prevent safety from being compromised. However, it is difficult to guarantee the performance of OOD detectors as it is difficult to characterize the OOD aspect of an instance, especially in high-dimensional unstructured data.

To distinguish between OOD data and data known to the learning component through the training process, an emerging technique is to incorporate variational autoencoders (VAE) within systems and apply classification or anomaly detection techniques on their latent spaces. The rationale for doing so is the reduction of the data domain size through the encoding process, which benefits real-time systems through decreased processing requirements, facilitates feature analysis for unstructured data and allows more explainable techniques to be implemented.

This study places probably approximately correct (PAC) based guarantees on OOD detection using the encoding process within VAEs to quantify image features and apply conformal constraints over them. This is used to bound the detection error on unfamiliar instances, ϵ, with user-defined confidence, $1 - \delta$. The approach used in this study is to empirically establish these bounds by sampling the latent probability distribution and evaluating the error with respect to the constraint violations that are encountered. The guarantee is then verified using data generated from CARLA, an open-source driving simulator.

Index Terms—Autoencoder, Conformal Prediction, Formal Verification, Generalized Error Bounds, Safety Guarantees

I. Introduction

Developments in artificial intelligence (AI) and machine learning (ML) have led to their implementations in safety-critical fields like transport, healthcare and security. Autonomous vehicles, amongst other cyber physical systems (CPS), use ML within their detection and decision-making subsystems. A reason for this is that ML models like deep neural networks (DNN) can create lower dimensional representations of abstract data that can be utilized for various tasks [9].

However, obstacles to widespread use are the lack of explainability regarding inner workings and the lack of guarantees on performance. The primary reason for this is the black-boxed nature of DNNs, which, due to the number of training parametres, make it difficult to provide safety assurances within the CPS context [19]. The necessity of these are highlighted by the fact that the performance estimation formed during the training/testing phase of development may

be different from the true performance of the system during deployment, oftentimes because of the existence of out-of-distribution (OOD) data that is unlikely to be present in the training phase [2].

OOD data refers to data that exist outside of the scope of data the model is familiar with. That is, instances that are out of the distribution defined by the dataset used during the training phase [11]. As it is impossible to account for all possible instances and states a system may encounter during the training phase, the system's behaviour toward OOD instances cannot be anticipated accurately and can be especially undesirable in safety-critical tasks [8]. For this reason, CPSs within safety-critical domains often contain subsystems dedicated to the detection and handling of OOD data [20].

There have been a number of studies that present implementations and frameworks for solutions to this problem, such as novelty or outlier detection and various OOD classifiers that have been developed using in-distribution benchmark datasets [11]. However, regardless of the algorithm used, an error-free OOD detection system is infeasible. Therefore, it is necessary to be able to guarantee the probability with which detection is conducted, especially in safety-critical tasks; i.e. to evaluate and bound the rate with which the subsystem fails to detect OOD instances.

An obstacle toward deriving general error bounds for OOD detection is the difficulty in characterizing the OOD aspect of high-dimensional data instances with respect to in-distribution properties [3]. That is, it is difficult to check if any properties of an instance are outside 'normal' parameters for high-dimensional data as the properties over which data is distributed, especially in image-based CPS, can be abstract [3]. As a result, creating definite, explainable constraints using in-distribution properties to evaluate whether instances are OOD, and thereby, bound the system's performance, is difficult.

A solution to this, utilized in systems described by [8], [20] and [7], is to use variational autoencoders (VAE) to parametrize the training data distribution with a fixed number of variables. VAEs are a class of DNNs that map high-dimension input data to lower-dimensional distributions that comprise latent spaces within the model. This results in the encoding of the distribution of training data to lower-dimension multivariate distributions. Studies have made use of this property in designing OOD detection systems by

attempting to equate OOD instances with outliers in the latent space [20] and constructing classifiers to define in-distribution safety constraints within this space [8].

As such, the objective of this study is to create a framework for guaranteeing and bounding OOD detection failure. The approach described in this study relies on the construction of constraints within the latent space that are used to define an in-distribution criteria for high-dimension data using the spatial coordinates of the encoding. Similar to [20], the constraints within this study are constructed using conformity-based classification based on a subset of known in-distribution data. Sampling the VAE latent distribution to find violations of these constraints allows for the construction of bounds on the error with which the system conducts OOD detection.

The guarantees on error provided in this study are through probably approximately correct (PAC) bounds. Two probabilistic measures are used to characterize the guarantee: the error level, $\epsilon \in (0, 1)$ and the confidence level, $\delta \in (0, 1)$. Through the sufficient sampling of the multivariate latent distribution, the approach provided in this study establishes that with a $1 - \delta$ confidence, the probability of OOD detection failure is less than ϵ. The correlation between sampling, constraint violation and confidence is used to bound the error probabilistically and provide an estimate of the performance of the system.

The structure of the remaining report will consist of a clarification of the assumptions made as well as the limitations of this study, the relevant works upon which this study is based, the theoretical approach taken and the results acquired from applying the theory.

II. RELATED WORKS

This investigation is related to two categories of research: *probably approximately correct guarantees for system safety* and *variational autoencoder based out-of-distribution detection.*

A. PAC-based Safety Guarantees

PAC learning was first introduced in [13] and has been utilized within several studies since. The objective of this framework is to be able to guarantee training within learning components to a certain extent with some confidence [5]. This has made the framework adaptable for formal verification purposes as it can be used to place error guarantees on certain properties of the output. This is cited as 'probably approximate safety verification' within [22].

There are a number of papers that address the specific topic using PAC-based guarantees to generalize error bounds within CPSs. Notable investigations in this area include [22], [19], [2] and [12]. Similar to the objective of this study, the PAC-based guarantees in the aforementioned works correlate the size of the training data to the failure rate with a particular level of confidence.

The error bounds for learning described in [5] and [13] use a generalized term correlated to the size of the hypothesis space to describe the target concept sample complexity. This is further generalized in [1] as a bound that is dependent on the VC-dimension of the model being used. The approach taken in [19] to place PAC guarantees applies this concept by attempting to estimate the VC-dimension of the classification algorithm.

In contrast to this, [22] proposes the formulation of PAC-based error bounds through the formulation of the problem as an optimization problem with the objective of minimizing the constraint violation probability. One of the main contributions of [22] is that stochastic perturbations within the input layer, with an underlying probability distribution, are factored into the derived error bounds. Because the problem investigated in this study can be framed similarly, a similar approach to [22] is utilized when deriving the error bounds.

However, because this study attempts to approximate safety constraints using conformal prediction [18] the guarantee placed on the constraints being accurate are incorporated into the PAC-based generalized error bounds for the entire system. To the best of our knowledge, aside from this paper, there are no existing studies on combining multiple types of guarantees when bounding the failure rate of an entire system.

B. VAE-based OOD Detection

Within recent years, several studies have emerged that use VAE latent encodings to reduce data dimensionality for tasks like classification [9] and anomaly detection [21]. [20] and [8] cite three clear benefits of doing so: firstly, the reduction in data dimensionality reduces the complexity of the required ML model, allowing more explainable techniques to be implemented [20]; secondly, the latent encoding allows for the quantification of high-dimensional features, increasing the robustness of classifiers applied in this space [8]; lastly, the reduction in dimensionality also reduces runtime [8].

There exist various extensions to techniques utilizing VAEs and this section is not exhaustive in detailing them. Instead, it will focus on the results of [8], [20] and [7], which make use of VAEs for the explicit purpose of OOD detection. [20] aims to train a VAE to construct a partially disentangled latent representation of a data set to be able to identify OOD data based on the targeted latent dimensions. A conformal predictor could then be used to determine a threshold value for OOD data along the tested dimension. [7] demonstrates a similar achievement using regression.

A research gap that should be noted is that none of the aforementioned studies place an emphasis on guaranteeing the OOD detection failure rate within this type of pipeline, which is necessary when designing safety-critical CPSs. Though, it is worth mentioning that while [8] and [20] build conformal predictors that operate with certain confidence within this space, there are no comments on the representation of the calibration/conformal set by the latent probability distribution, which is required to bound the failure rate of the entire OOD detection system, including the VAE's encoding. To the best of our knowledge, aside from this paper, there are no existing studies on this.

III. PRELIMINARIES AND DEFINITIONS

The investigation conducted in this paper is dependent on various existing techniques and the results from studies that have been conducted in the past. This section will provide background knowledge and define related terminology that will be used.

A. Safety Constraints

The safety verification procedure in this study is conducted by verifying that sampled data instances lie within a 'safe' region of the encoded hyperspace. This region is defined using a set of safety constraints, denoted by S_n for n constraints.

In previous studies like [22], [4] and [10], assuming the dimensionality of the instance is k, the safe region, $\mathbb{S} \subseteq \mathbb{R}^k$, is equivalent to the set of values defined by x.

$$\mathbb{S} = \left(x \in \mathbb{R}^k \mid \max_{j=1,2...n} S_j(x) \le 0 \right) \tag{1}$$

Though the safety constraints implemented in this study are based on the Inductive Conformal Prediction framework (ICP) of a classification method discussed in [18], they are adapted using (1) as a basis.

B. Inductive Conformal Prediction

ICP is a framework for prediction that relies on the degree to which future instances conform with known data and accordingly issues a guarantee on the confidence of the prediction.

For ease of notation, the space, Z, created by the Cartesian product of the feature space and label space, X and Y respectively, encompasses the training set $Z_M := (z_1...z_M)$, with training samples $z_i = (x_i, y_i) \in Z_M$. The training set can be split into two sets, Z_L, the training set, and Z_{M-L}, the calibration set, with $L < M$ [17].

The role of the calibration set is made apparent when considering the conformity measure, C, a function that outputs a value in proportion to the degree with which future data samples conform to the calibration set. [18], where conformal prediction was introduced, establishes C as a function that compares the output of a predictor f with the label for a data instance. Δ is used to denote the comparator.

$$C(Z_{M-L}, z_i) := \Delta(y_i, f(x_i)) \tag{2}$$

Within the context of a classification problem, C is used to evaluate the conformity score of a data instance, x, with all labels $y \in Y$ and outputs a set prediction for the potential class label of the instance based on the conformity of the label-instance pair with the existing calibration set.

Given a significance level, $\beta \in (0,1)$, where $1 - \beta$ is the confidence level, a threshold, t^*, can be calculated from the calibration set by ordering the conformity scores of the calibration set and taking the β percentile score. Letting T be the ordered set of sorted conformity scores from the calibration set, the following is defined.

$$T := Sorted\left(t_i | t_i = \max_{y_i \in Y} \left(\Delta(y_i, f(x_i)) \right), (x_i, y_i) \in Z_{M-L} \right)$$
$$with \ (t_i \le t_{i+1}),$$
$$t^* = t_{\lfloor \beta(M-L) \rfloor} \tag{3}$$

Note that it is assumed that the degree of conformity is greater for larger values of t^*. Using t^*, a set prediction can be constructed for a data instance, x.

$$\Gamma(Z_{M-L}, x) = \{y | \Delta(y, f(x)) > t^*\} \tag{4}$$

The set predictor will have made an error if the correct label is not an element of the prediction. The probability of this occurring for a given prediction is less than β and, therefore, there is a more than $1 - \beta$ confidence that the set predictor is correct [18]. It is also worth noting that if none of the labels conform to a data instance, the predictor yields the null set, $\varnothing$. This property is useful for predicting OOD instances.

In order for ICP to hold, the following assumption has to be made [17].

Assumption 1: The elements of the calibration set are exchangeable.

This implies that the distribution of the calibration set is to be representative of the distribution of the training set. Under these circumstances, the set prediction for a given data instance is invariant to different combinations of the calibration set and the prediction error is held less than β.

C. Variational Autoencoders

VAEs are a class of generative DNNs based on the encode-decode approach of autoencoders. The model assumes the existence of an underlying prior probability distribution of the training data. The model approximates the prior using a multivariate Gaussian distribution of fixed dimensions that comprises the latent dimension [6]. The resulting trained distribution is representative of the distribution of data within the latent space. Sampling from this distribution produces new data that preserves the learnt characteristics of the dataset.

If a VAE is trained using in-distribution data, it stands to reason that the probability density function used to represent the latent distribution corresponds with the degree to which instances in the latent space are in-distribution. However, it is difficult to place a guarantee on the robustness of the encoding as well as the degree to which the latent distribution represents the in-distribution characteristics of the training set. As such, this study, similar to [8] and [20], utilizes encodings of known in-distribution data to create safety constraints within this space.

D. Probably Approximately Correct Guarantees

Prior to formulating the safety verification problem and proposing a solution, it is necessary to describe the type of probabilistic guarantee that will be used.

PAC learning is a concept that describes the efficient learning of a target hypothesis through approximation. Formally, the efficiency ascribed to PAC is in the form of a probably approximate learning guarantee, i.e. with at least $1 - \delta$ probability, $\delta \in (0, 1)$, the learnt concept will approximate the target concept with greater than $1 - \epsilon$ accuracy, given $\epsilon \in (0, 1)$.

The motivation behind using a probabilistic, $1 - \epsilon$ approximation of the target concept is to reduce sample complexity, formally stated in [5] using the following equation, where N denotes the sample complexity and H_N denotes the size of the hypothesis space.

$$N \geq \frac{1}{\epsilon} \left(\ln(H_N) + \ln \left(\frac{1}{\delta} \right) \right) \tag{5}$$

Inferring from inequality (5), any subsequent increase in either confidence or accuracy requires a larger increase in sample size. Therefore, in a system where time is a constraint to learning, an optimal level of accuracy can be guaranteed with a reduction in the samples used for training.

The idea of a probably approximate safety guarantee, or PAC barrier certificate in [22], is borrowed from this concept. The application of the learning theory as a safety guarantee is to be able to, similarly, verify that with more than $1 - \delta$ confidence, the safety constraints are violated with less than an ϵ-level probability using a fixed sample size, N.

To appropriate this framework, Assumption 2, the invariance assumption, and Assumption 3, learnable regularity, are required [14].

Assumption 2: The distribution of data samples that are utilized during the training process is invariant to the distribution of the source of the samples.

This assumption is required to make inferences about the error bound for the system's performance during deployment.

Assumption 3: There exist regularities in the data that can be used to efficiently categorize and learn the target concept feasibly.

This assumption is necessary to evaluate OOD detection error. A step in doing so is to encode the set of learnable characteristics within the latent dimensions of the VAE architecture as well as categorize in-distribution data through these using arbitrary safety constraints.

Based on the PAC framework, the problem formulation for this study is presented in Problem 1.

Problem 1: *Given an out-of-distribution detection system consisting of a trained variational autoencoder, safety constraints identifying in-distribution characteristics within the latent encoding and a confidence level $\delta \in (0, 1)$, derive bounds $\epsilon \in (0, 1)$ such that the detection system, with at least $1 - \delta$ confidence, misidentifies OOD instances as in-distribution with less than ϵ probability.*

The objective of Problem 1 is to compute ϵ, the OOD detection error represented by the false positive in-distribution rate, with a particular confidence. However, this error can be minimized by categorizing all instances encountered as OOD. Therefore, a trade-off between in-distribution detection accuracy and OOD detection accuracy is inevitable. Intuitively, Problem 1 can be restated as the computation of the maximum tolerable OOD detection error for the system.

IV. PROBLEM FORMULATION

This section formalizes Problem 1 as an optimization problem and introduces the theorems necessary to provide a solution.

The intuitive formulation from Problem 1 is the calculation of the dissociation of $\mathbb{S}$, the space defined by the safety constraints, with the OOD regions and distributions of data outside of the specified in-distribution. Thereby, the problem would be the computation of ϵ, for the following inequality (6), given that θ represents the VAE's learnt distribution over the latent space, given a data instance, x, and given n safety constraints.

$$P \left(x \notin \theta \mid \max_{j=1,2...n} S_j(x) \leq 0 \right) \leq \epsilon \tag{6}$$

With the satisfaction of the safety constraints, if an instance is OOD with θ, then the constraints are in error. The difficulty with the computation of the LHS is that learning the corresponding distribution of OOD data within the latent space and sampling from it is infeasible. An alternative formulation would be (7), which describes the probability of an instance belonging to θ given that the safety constraints are satisfied. The corresponding probability is an upper bound for $1 - \epsilon$.

$$P \left(x \in \theta \mid \max_{j=1,2...n} S_j(x) \leq 0 \right) \geq 1 - \epsilon \tag{7}$$

A potential solution to this is the integral of the latent multivariate Gaussian distribution defined within $\mathbb{S}$. However, given the probabilistic constraints that have been constructed, a more appropriate formulation would be as the following chance constrained optimization problem (CCP), where U is a user defined upper bound, in similar vein to [22].

$$\min_{\lambda \in \mathbb{R}} \lambda \ s.t.,$$
$$P \left(x \in \theta \mid \max_{j=1,2...n} S_j(x) \leq \lambda \right) \geq 1 - \epsilon,$$
$$0 \leq \lambda \leq U \tag{8}$$

CCPs are computationally hard problems. However, the results of [15] and [16] show they can be relaxed at the cost of the robustness of the solution using a scenario optimization approach. That is, the solution to a deterministic relaxation of the original problem is a valid solution to the original problem with a guaranteed confidence. For this reason, the objective of the problem is to minimize λ with respect to the constraints, allowing for reasonable deviation from the original problem. The relaxation of (8) given this approach is (9).

$$\min_{\lambda \in \mathbb{R}} \lambda \ s.t.,$$

$$for \ each \ i \in \{1, 2, 3...N\} \ ,$$

$$\max_{j=1,2...n} S_j(x_i) - \lambda \leq 0,$$

$$0 \leq \lambda \leq U \quad (9)$$

In doing so, the chance-based constraints are replaced by N instantiations of the constraints that can be violated with an ϵ probability by feasible solutions to the problem. The confidence with which this is applicable is described by Theorem 1, presented in [15].

Theorem 1 [15]: *Given a value $\delta \in (0, 1)$, if ϵ, N and r are such that the following condition holds,*

$$\binom{r+d-1}{r} \sum_{i=0}^{r+d-1} \binom{N}{i} e^i (1-e)^{N-i} \leq \delta \quad (10)$$

with d being the number of optimization variables and $\mathbb{P}^N$ being the N-fold probability of constraint satisfaction with an ϵ error level, the following also holds

$$\mathbb{P}^N \left(P \left(x \in \theta \mid \max_{j=1,2...n} S_j(x) \leq 0 \right) \geq 1 - \epsilon \right) \geq 1 - \delta \quad (11)$$

Theorem 1 establishes a relation between the number of samples drawn, N, the number of constraint violations, r, the tolerable violability of the constraints, ϵ, and the confidence with which this occurs, δ, for a fixed number of optimization variables, d. Based on this this, if the condition in (10) is met, the safety constraints will be satisfied with an ϵ error level with a confidence of $1 - \delta$ [15].

Applying Theorem 1 directly to the problem described in (9) reduces the amount of computation required by loosening the constraints and increasing the size of the feasible region within $\mathbb{R}^k$. Permitting violations of the safety constraints for N sampled instances to an ϵ degree guarantees the solution with a $1 - \delta$ confidence assuming Theorem 1 holds, as applied in [22].

However, the formulation of Problem 1 is regarding the derivation of ϵ given δ rather than the converse. The following section describes how Theorem 1 can be used to do so.

V. Guaranteeing Out-Of-Distribution Detection

A. Constructing Safety Constraints

The difficulty with placing safety constraints on the latent space exists because the encoding and decoding processes are opaque and to verify that the output is safe based on the sampled latent variables requires being able to map the latent space to the output space. Similarly, determining the latent variables that correspond to the right generative factor and their correlation is an exhaustive process.

Some similarities can be drawn between the notion of conformity within the ICP framework and conventional safety constraints for OOD detection problems; i.e. data instances that fail to meet certain criteria specified by the safety constraints vs. the more abstract comparison of the conformity measure with the calibration set threshold.

A potential conformity metric used to create a score for a data instance is the probability density of the calibration set at the location of the instance in the latent space of the VAE [8]. If the density exceeds the threshold, it is probable that the instance conforms. This can be tested for each label when forming the set prediction.

A unique property of the set predictor that can be utilized to identify OOD instances is the null set prediction, $\varnothing$, which implies a lack of conformity with any class [18]. However, it should be noted that the expected rate of exclusion of the correct label from the set prediction is β and it follows that the confidence of the predictor in constructing the correct prediction is $1 - \beta$. Therefore, when the set predictor outputs $\varnothing$, it does so with a $1 - \beta$ confidence. Using the null set prediction property of ICP as the definition for OOD instances, the safe region can be constructed using the following.

$$\mathbb{S} = \left\{ x \in \mathbb{R}^k \mid \max_{j=1,2...n} \left(C(Z_{M-L}, (x, y_j)) \right) \geq t^* \right\} \quad (12)$$

That is, for a calibrated set predictor,

$$\mathbb{S} = \left\{ x \in \mathbb{R}^k \mid \Gamma \left(Z_{M-L}, x \right) \neq \varnothing \right\} \quad (13)$$

Implementations of the density based conformity metric consist of approximations like kernel density estimation and K-nearest neighbor distance scores.

Unlike the construction of safety constraints using conventional classifiers like support vector machines in [8], the conformity based safety constraints are defined using a calibration set as well as a confidence measure, β, that dictates the degree of deviation permitted for a new data instance from the characteristics of the calibration set. In turn, this allows for more flexibility as well as minimal representation error when determining safety constraints when compared with the method in [8] that requires the support vector algorithm to do so. For a visual comparison, refer to Figures 1 and 2.

Furthermore, in order to construct the constraints using support vectors, a number of known OOD samples must be included in the training set. This restriction is eliminated when using a conformal predictor.

The following algorithm can be used to construct the safety constraints over the training set.

Algorithm 1 : *Establishing Safety Constraints*

Pre-conditions :
- Trained VAE, latent distribution θ over $\mathbb{R}^k$,
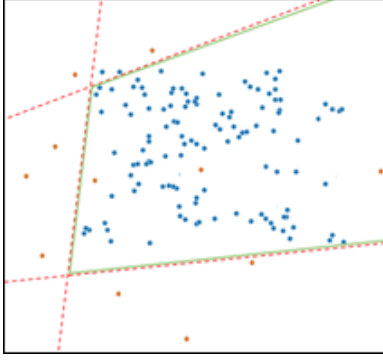- Let $Z_{M-L} \subseteq \mathbb{S}$ be the calibration set, $M, L \in \mathbb{Z}^+$,

Fig. 1. Safe region created by support vectors highlighted by green.

Fig. 2. Safe regions created using a uniform kernel estimation at different significance levels (β).

- Significance $\beta \in (0, 1)$;

Procedure :

1) Construct T, the set of conformity scores for each element within the calibration set using the kernel density estimate (KDE) for each point $z \in Z_{M-L}$;
2) Sort T in ascending order, $t_i < t_{i+1}, t_i \in T$;
3) Establish the threshold $t^* = t_{\lfloor \beta(M-L) \rfloor}$;
4) Establish the constraints by building set predictor Γ,
   a) i.e. $x \in \theta$ is an element of the safe region iff. $\Gamma(Z_{M-L}, x) \neq \varnothing$;

Though Algorithm 1 utilizes the KDE algorithm within this study, note that T can be established using any density based metric with a similar ordering property.

B. Deriving Error Bounds for OOD Detection

Assuming ideal conditions are met and there exists a solution to (9) where $\lambda \leq 0$, that is, an absence of constraint violations are encountered, the following holds [15].

$$\epsilon \geq 1 - \delta^{1/N} \tag{14}$$

This can be derived from Theorem 1 by making assumptions regarding the number of violated constraints and, intuitively, defines the relation between ϵ and δ as δ is an N-fold probability and is defined in (11). However, in the instance that λ is greater than zero, (14) does not hold. For this, [22] applies Chernoff bounds to the binomial condition in Theorem 1, (10), and, through inequality (8) in [15], presents adjusted error bounds that account for solutions where constraints are violated.

$$\epsilon \geq \min\left\{1, \frac{1}{N}\left(r + \ln\frac{1}{\delta} + \sqrt{\ln^2\frac{1}{\delta} + 2r\ln\frac{1}{\delta}}\right)\right\} \tag{15}$$

For the specific application of (15) within this study, the error bound presented in (15) can be tightened further conside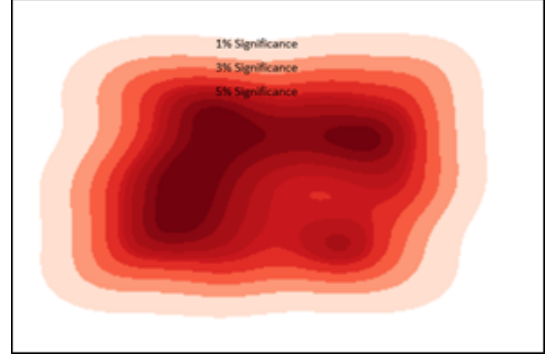ring the confidence parameter, β, from Section VI-A that describes the probability with which a true constraint violation has taken place assuming that a sampled instance does not conform to the calibration set. The adjustment is to the number of detected constraint violations by a factor of $1 - \beta$, the confidence of the conformal prediction. The proof for (15) and adjustment made to it in (16) is detailed in section VIII.

$$\epsilon \geq \min\left\{1, \frac{1}{N}\left(r(1-\beta) + \ln\frac{1}{\delta} + \sqrt{\ln^2\frac{1}{\delta} + 2r(1-\beta)\ln\frac{1}{\delta}}\right)\right\} \tag{16}$$

With this, the algorithm required to conduct the safety verification of the OOD detection system and bound the performance is a counting algorithm that records the number of constraint violations, or forced relaxations, within N sampled instances and bounds the number of potential future violations as N approaches infinity with $1 - \delta$ confidence. Algorithm 2 describes this process with greater detail.

Algorithm 2 : *Computing* ϵ

Pre-conditions :
- Trained VAE, latent distribution θ over $\mathbb{R}^k$,
- Set predictor Γ,
- Calibration set $Z_{M-L} \subseteq \mathbb{R}^k$,
- Significance $\beta \in (0, 1)$,
- Instantiated values $N \in \mathbb{Z}^+$ and $\delta \in (0, 1)$;

Procedure :

1) Initialize variable r to 0;
2) Loop N times,
   a) Generate sample x from θ,
   b) If $\Gamma(Z_{M-L}, x) = \varnothing$, increment r;
3) If $1 < \frac{1}{N}\left(r(1-\beta) + \ln\frac{1}{\delta} + \sqrt{\ln^2\frac{1}{\delta} + 2r(1-\beta)\ln\frac{1}{\delta}}\right)$, return 1,
   a) Else, return
      $\frac{1}{N}\left(r(1-\beta) + \ln\frac{1}{\delta} + \sqrt{\ln^2\frac{1}{\delta} + 2r(1-\beta)\ln\frac{1}{\delta}}\right)$;
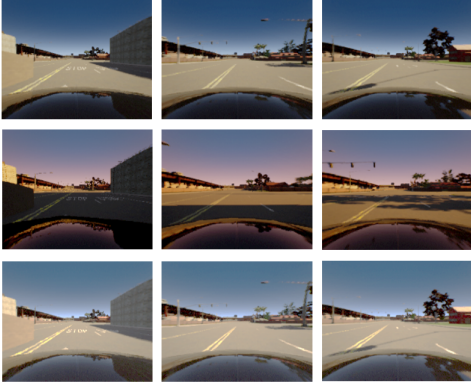

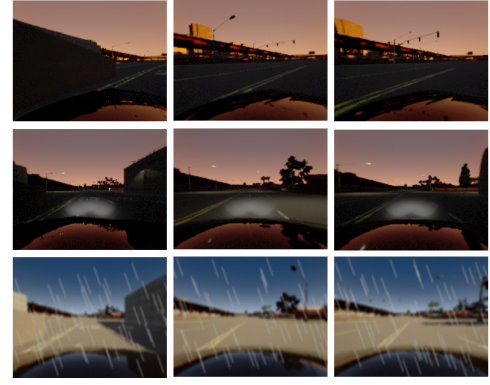Fig. 3. In-Distribution CARLA Simulation Data

Fig. 4. Out-of-Distribution CARLA Simulation Data

VI. EVALUATION OF BOUNDS

This section describes the results of applying the theories in Section V. All computations were performed using a Google Colab environment with 12GB memory, 100GB disk space, 2.3GHz CPU and a Tesla k80 GPU.

A. VAE Properties

The architecture of the VAE in this experiment is as follows.

The model is divided into the encoder and decoder. Within the encoder, there are five layers of convolution, followed by five densely connected layers. Within the decoder, there are four densely connected layers followed by four layers of convolution. The latent encoding is comprised of 16 variables as it was assumed that this value was an appropriate upper bound for the number of generative factors for the DVM-CAR dataset. Lastly, through a grid search, the hyperparameter coefficient of the KL-divergence term in the loss function was set to 2.2.

B. Data Properties

In order to verify that the method described in this paper can be successfully applied to the OOD detection system described, the dataset used to train the VAE consisted of images generated from running the CARLA driving simulator within fixed environmental parameters, e.g. rain, sunlight and location. The motivations behind using this dataset are because it is highly controllable with easily quantifiable OOD instances and the driving simulator is representative of the data complexity encountered during deployment.

All images in the training dataset contain similar properties and are in-distribution. The partition between in-distribution data and OOD data is through the following features that were set upon running the simulator:

- Precipitation - Any amount is OOD;
- Brightness - Any value under 0.5 is OOD;
- Road Segment - Any segment apart from the one shown in Fig. 3 is OOD.

Samples from the simulation that are known to be OOD are indicated in Fig. 4.

1600 in-distribution images are used in the VAE training process, from which 200 are selected for the calibration set.

C. Computing Error Bounds

In this experiment, the safety constraints of the form of Equation (13) are computed using Algorithm 1 with a subset of training samples that comprise the calibration set. The objective of Algorithm 1 is to establish a metric by which future samples can be scored to verify conformity with the training set. If the degree of conformity fails to exceed a threshold, t^*, the instance is OOD. Spatially, this is equivalent to partitioning the encoded $\mathbb{R}^k$ space into a safe and unsafe region determined by the density of the encoded calibration set within that region.

During the computation of the safety constraints, sets of 200 in-distribution CARLA samples were used to construct the calibration set. These samples were encoded and the KDE algorithm was applied with a uniform kernel to establish the density of the calibration set at each point. These values were then normalized and ordered. For all subsequent experiments, the significance level, β, taken was 0.0275.

Based on this and Equation (3), the 5th element of the ordered set of conformity scores was used as the threshold for OOD prediction. The following table describes the trials that took place using the conformity predictor at various levels of confidence and for various sample sizes.

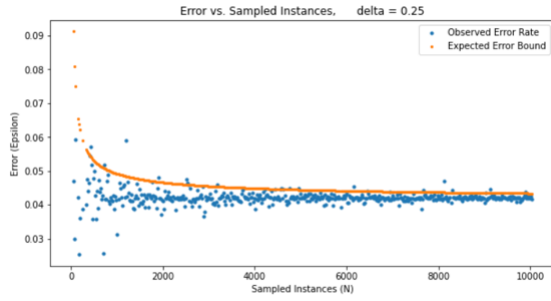| Table 1. Sample Experiment Results				
N	δ	r	$\frac{r}{N}$	ϵ
10^2	10^{-6}	5	0.0500	0.4141
10^3	10^{-6}	45	0.0450	0.1009
10^4	10^{-6}	436	0.0436	0.0559
10^5	10^{-6}	4301	0.0430	0.0457
10^6	10^{-6}	43235	0.0432	0.0433
10^5	10^{-4}	4311	0.0431	0.0449
10^5	10^{-6}	4359	0.0436	0.0460
10^5	10^{-8}	4283	0.0428	0.0458
10^5	10^{-10}	4277	0.0428	0.0463

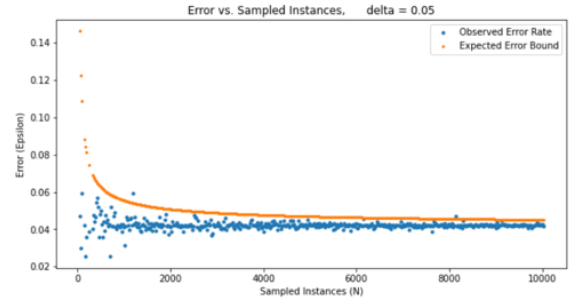Fig. 5 . Observed Error and Error Bound vs. Sample Size for $\delta = 0.25$

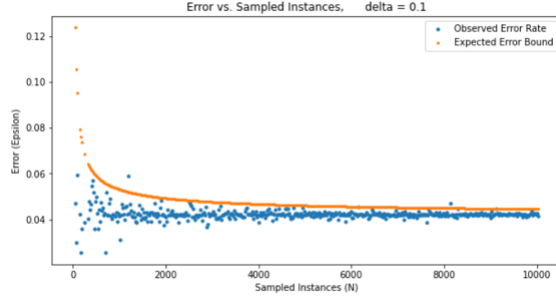Fig. 7 . Observed Error and Error Bound vs. Sample Size for $\delta = 0.05$

Fig. 6 . Observed Error and Error Bound vs. Sample Size for $\delta = 0.1$
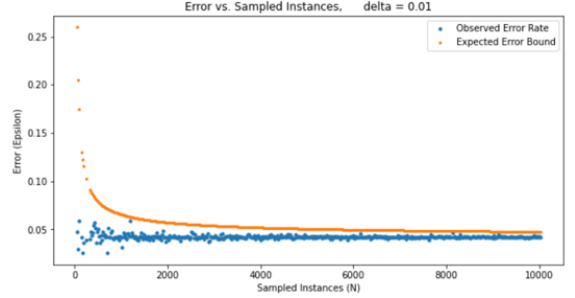

Fig. 8 . Observed Error and Error Bound vs. Sample Size for $\delta = 0.01$

Table 1 describes the results from the experiment and the performance of Algorithm 2 in establishing error bounds. Given the sample size, N, and confidence, δ, Algorithm 2 computes the number of samples in violation of the established safety constraints, r, yielding the proportion of samples in violation of the constraints, $\frac{r}{N}$. From these values, Equation (16) could be used to bound the error rate, ϵ.

The aim of this study is to be able to provide an upper bound for the value $\frac{r}{N}$, ϵ, which must be able to bound $\frac{r}{N}$ with $1 - \delta$ confidence. As such, there are two assessments to be made regarding the bounds that have been derived in this study, namely, the validity and effectiveness of the bounds. The validity of the bounds is the determination of whether or not ϵ is greater than $\frac{r}{N}$. The effectiveness is the measure of the tightness of the bound, i.e. how distant ϵ is from $\frac{r}{N}$.

Based on the figures observed during the experiment, the number of data instances in violation of the constraints are approximately 4% of the sample size. The validity can be established by observing that the error bound is greater than the proportion of constraint violations relative to the number of instances sampled with greater than $1 - \delta$ confidence. Though the bound is greater than the observed error rate for all recorded trials in Table 1, the effects of adjustments in the confidence parameter to the error bound are more explicit in Fig. 5-8.

Additional observations from the table are:

- that ϵ increases as δ decreases, widening the bounds to increase the probability that the true error rate has, in fact, been bounded;
- that ϵ decreases as N increases, tightening the bounds as

the latent distribution is sampled further and the sampled error rate approaches the expected error rate.

Based on these observations, the tightness of the bounds is also made clear as, for a fixed confidence, as the number of instances sampled increases, ϵ tends infinitely close to $\frac{r}{N}$ while still providing an upper bound, as observed in row 5 of Table 1.

Fig. 5-8 indicate the relation between error and the number of instances sampled for fixed values of δ. The points on the graph indicate the observed error rate within a trial and the expected error bound, ϵ, that places an upper bound guarantee on the error rate with $1 - \delta$ probability. A violation of the error bound during a trial can be expected in $\frac{1}{\delta}$ trials. As such, the experiments depicted in Fig. 5-8 are of trials where δ is in the range $[0.25, 0.1, 0.05, 0.01]$ and the number of data points recorded for each confidence value are 500. The percentage of trials where the observed error rate exceeds the expected error bound is denoted in Table 2 alongside the corresponding graph and the confidence.

Table 2. Error Bound Violations		
Fig.	δ	$P(\frac{r}{N} > \epsilon)$
5	0.25	0.038
6	0.10	0.008
7	0.05	0.004
8	0.01	0.000

The graphs validate the relations inferred from the observations noted in the table and validate the error bound derived in this study with a greater than $1 - \delta$ probability.

VII. Conclusions

This study successfully derives guarantees for OOD detection with fixed levels of confidence that are within sampled error bounds for uncertain safety constraints. The framework for doing so utilizes VAEs to quantify the features comprising the distribution of training data and placing ICP-based safety constraints based on samples that conform to the in-distribution label. The algorithm for the error bound calculation depends on sampling from the VAE's learnt distribution over the latent dimension and counting the samples in violation of the constraints. Lastly, testing with a dataset of images from the CARLA driving simulator proved that the derived bounds are valid for all error-confidence pairs.

The results of this study present a framework to predict system performance prior to deployment and independent of the type of safety constraints. And while an implementation of the technique on the CARLA driving simulator demonstrates its practicality, it raises questions from a theoretical standpoint as to developments that could be made in future studies. Extensions to this study should consider the effects that varying the sample size of the calibration set has on the error bound derivation as it is used to construct safety constraints that approximate the in-distribution characteristics being assessed. This may become a consideration for real-time OOD detection implementations that require smaller calibration sets to ensure runtime feasibility.

We hope that the results described in this paper demonstrate the reliability of PAC-based formal verification and inform future studies that aim to guarantee CPS safety prior to deployment.

VIII. Appendix : Proof of Inequations (15) and (16)

The derivation of (15) is absent in [22], from where this paper cites it. As such, this paper attempts to re-derive this bound within this section, beginning with inequation (8) from [15], restated in (17), and the associated preliminaries.

$$r \leq \epsilon N - d + 1 - \sqrt{2\epsilon N \ln \frac{(\epsilon N)^{d-1}}{\delta}} \qquad (17)$$

Implying the following,

$$\epsilon \geq \frac{1}{N} \left(r + d - 1 + \sqrt{2\epsilon N \ln \frac{(\epsilon N)^{d-1}}{\delta}} \right) \qquad (18)$$

An assumption being made in [15] is that $\epsilon N \geq r + d - 1$. Substituting into (17),

$$\epsilon \geq \frac{1}{N} \left(r + d - 1 + \sqrt{2(r+d-1) \ln \frac{(\epsilon N)^{d-1}}{\delta}} \right) \qquad (19)$$

Following from the previous assumption is (19).

$$\epsilon N - r \geq d - 1 \qquad (20)$$

However, the definition of the expected value of r, given a sample size of N, is as follows,

$$E_N[r] = \epsilon N \qquad (21)$$

Furthermore, given the Law of Large Numbers,

$$\lim_{N \to \infty} r - \epsilon N = 0 \qquad (22)$$

Thus, substituting (22) into (20), $d \leq 1$. The substitution of the resultant into (19) ensures that (23) holds.

$$\epsilon \geq \frac{1}{N} \left(r + d - 1 + \sqrt{2(r+d-1) \ln \frac{1}{\delta}} \right) \qquad (23)$$

Thus,

$$\epsilon \geq \frac{1}{N} \left(r + d - 1 + \sqrt{2r \ln \frac{1}{\delta} + 2(d-1) \ln \frac{1}{\delta}} \right) \qquad (24)$$

As $\ln \frac{1}{\delta} > 2(d-1)$ for the following range, $0 < \delta << 1$, the following substitution can be made to upper bound the RHS expression in (23).

$$\epsilon \geq \frac{1}{N} \left(r + \ln \frac{1}{\delta} + \sqrt{2r \ln \frac{1}{\delta} + \ln \frac{1}{\delta} \ln \frac{1}{\delta}} \right) \qquad (25)$$

(25) is equivalent to (15).

(16) is produced by asserting Assumption 1 as well as the proposition in [17] regarding the validity of predictions made using conformal predictors, that conformal set predictions are made with $1 - \beta$ confidence. Therefore, the bound on ϵ in (15) can be further tightened using the expected value of the erroneous null set detections.

Of the r recorded constraint violations, it is expected that there are βr errors and, therefore, the value of r used within (15) can be adjusted by a factor of $1 - \beta$ to derive (16), restated in (26) with the adjusted error bound ϵ^*.

$$\epsilon^* \geq \min \left\{ 1, \frac{1}{N} \left(r(1-\beta) + \ln \frac{1}{\delta} + \sqrt{\ln^2 \frac{1}{\delta} + 2r(1-\beta) \ln \frac{1}{\delta}} \right) \right\} \qquad (26)$$

Note that $\epsilon^* \leq \epsilon$ because the RHS of (26) is less than the RHS of (25) for $\beta \in [0, 1]$. Therefore, while (26) is being used to substitute (25) within this study, using (25) is a valid approach to determining the OOD detection failure rate as it presents an upper bound for (26).

IX. Acknowledgements

REFERENCES

[1] A. Ehrenfeucht, D. Haussler, M. Kearns and L. Valiant, "A general lower bound on the number of examples needed for learning", Information and Computation, Vol. 82-3, pp. 247-261, September 1989.

[2] A. Farid, D. Snyder and A. Ren, "Failure Prediction with Statistical Guarantees for Vision-Based Robot Control", Proceedings of Robotics: Science and Systems, June 2022

[3] A. Pereira and C. Thomas, "Challenges of Machine Learning Applied to Safety-Critical Cyber-Physical Systems", Machine Learning and Knowledge Extraction, November 2020

[4] C. Dawson, Z. Qin, S. Gao and C. Fan, "Safe Nonlinear Control Using Robust Neural Lyapunov-Barrier Functions", 5th Annual Conference on Robot Learning, June 2021.

[5] D. Haussler, "Probably Approximately Correct Learning", AAAI-90 Proceedings, Vol. 2, pp. 1101-1108, July 1990.

[6] D. Kingma and M. Welling, "Autoencoding Variational Bayes", International Conference on Learning Representations, April 2014.

[7] F. Cai, J. Li and X. Koutsoukos, "Detecting Adversarial Examples in Learning-Enabled Cyber-Physical Systems using Variational Autoencoder for Regression", In IEEE Security and Privacy Workshops, May 2020.

[8] F. Cai and X. Koutsoukos, "Real-time Out-of-distribution Detection in Learning-Enabled Cyber-Physical Systems", 11th International Conference on Cyber-Physical Systems, April 2020.

[9] H. Li, H. Wang, Z. Yang and M. Odagaki, "Variation Autoencoder Based Network Representation Learning for Classification", Proceedings of ACL 2017, January 2017.

[10] H. Zhao et. al., "Synthesizing barrier certificates using neural networks", Proceedings of the 23rd International Conference on Hybrid Systems: Computation and Control, April 2020.

[11] J. Yang, X. Zeng, T. Chen and Z. Liu, "Generalized Out-of-Distribution Detection: A Survey", Available on arXiv: Computer Vision and Pattern Recognition, October 2021.

[12] L. F. O. Chamon and A. Ribeiro, "Probably Approximately Correct Constrained Learning", Proceedings of the 34th International Conference on Neural Information Processing Systems, December 2020.

[13] L. Valiant, "A theory of the Learnable", Communications of the ACM, Vol. 27-11, pp. 1134-1142, November 1984.

[14] L. Valiant, "Probably Approximately Correct", 978-0465032716.

[15] M. Campi and S. Garatti, "A Sampling-and-Discarding Approach to Chance-Constrained Optimization: Feasibility and Optimality", Journal of Optimization Theory and Applications, Vol. 148-2, pp. 257-280, February 2011.

[16] M. Campi and S. Garatti, "The Exact Feasibility of Randomized Solutions of Uncertain Convex Programs", SIAM Journal on Optimization, Vol. 19-3, January 2008.

[17] V. Vovk, "Conditional Validity of Inductive Conformal Predictors", Machine Learning, Vol. 92, pp. 349-376, September 2012.

[18] V. Vovk, A. Gammerman and G. Shafer, "Algorithmic Learning in a Random World", 978-0-387-25061-8.

[19] S. Park, O. Bastani, N. Matni and I. Lee, "PAC Confidence Sets for Deep Neural Networks via Calibrated Prediction", International Conference on Learning Representations 2020, February 2020.

[20] S. Ramakrishna, Z. Rahiminasab, G. Karsai, A. Easwaran and A. Dubey, "Efficient Out-of-Distribution Detection Using Latent Space of β-VAE for Cyber-Physical Systems", ACM Transactions on Cyber-Physical Systems, Vol. 6-2, pp. 1–34, April 2022.

[21] S. Suh, "Echo-state conditional variational autoencoder for anomaly detection", 2016 International Joint Conference on Neural Networks, July 2016.

[22] X. Bai, M. Franzle, H. Zhao, N. Zhan and A. Easwaran, "Probably Approximate Safety Verification of Hybrid Dynamical Systems", Proceedings of 21st International Conference on Formal Engineering Methods, August 2019.