IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 22, 2022, accepted April 13, 2022, date of publication April 18, 2022, date of current version April 26, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3168003

A Generative Approach to Open Set Recognition Using Distance-Based Probabilistic Anomaly Augmentation

JOEL GOODMAN[1], (Senior Member, IEEE), SHAHRAM SARKANI[2], AND THOMAS MAZZUCHI[2]

[1]U.S. Naval Research Laboratory, Washington, DC 20375, USA
[2]School of Engineering and Applied Science, The George Washington University, Washington, DC 20052, USA

Corresponding author: Joel Goodman (joel.goodman@nrl.navy.mil)

This work was supported by the Office of Naval Research (ONR), Arlington, VA, USA.

ABSTRACT Machine learning (ML) algorithms that are used in decision support (DS) and autonomous systems commonly train on labeled categorical samples from a closed set. This, however, poses a problem for deployed DS and autonomous systems when they encounter an anomalous pattern that did not originate from the closed set distribution used for training. In this case, the ML algorithm that was trained only on closed set samples may erroneously identify an anomalous pattern as having originated from one of the categories in the closed set, sometimes with very high confidence. In this paper, we consider the problem of unknown pattern recognition from a generative perspective in which additional synthetic training samples that represent anomalies are added to the training data. These synthetic samples are generated to optimally balance the desire to place anomalies all along the boundary of the training set in feature space, while not adversely effecting core classification performance on the test set. We demonstrate the efficacy of distance-based probabilistic anomaly augmentation (DPAA) that is proposed in this paper for a diverse set of applications such as character recognition and intrusion detection, and compare its combined classification and identification performance to both recent open set and more traditional novelty detection approaches.

INDEX TERMS Machine learning, outlier and novelty detection, open set recognition, anomalies, generative and discriminative architectures.

I. INTRODUCTION

Novelty and outlier detection are popular approaches for recognizing anomalies and/or anomalous behavior and are commonly used in decision support and autonomous applications such as medical diagnostics, fault detection in manufacturing processes, fraud and intrusion detection [1]–[4]. The objective of novelty detection is to identify patterns that are not representative of the data used to train the detector, and can broadly being characterized as either being discriminative or generative in nature [5]. Discriminative approaches can further be categorized as being statistical or distance-based [6]. Popular examples of distance-based approaches include the one class support vector machine (OC-SVM) [7], isolation forests (IF) [8], and the deep neural network (DNN) autoencoder (AE) [9] each of which uses a distance measure from the training set boundary to identify anomalies. An OC-SVM identifies support vectors that encapsulates the training set, and rejects any new sample that falls on the other side of the boundary [10]. Autoencoders compress the data down into a lower dimensional latent space, and measures the (typically 2-norm) reconstruction error after decompression to determine if the sample is an anomaly [11]. Isolation forest is a tree ensemble based method that measures anomalies by the depth of decision before reaching a leaf node. The more shallow and the fewer the cuts needed to reach a leaf node, the more likely it is that the sample is an anomaly [12]. Although OC-SVMs, AEs and Isolation Forests use different distance measures to identify anomalies, they represent some of the most mature and commonly employed discriminative approaches to novelty detection [13].

The associate editor coordinating the review of this manuscript and approving it for publication was Seifedine Kadry.

Statistical approaches to anomaly detection, in contrast to distance-based methods, try to model the distribution of the training samples from the known categories and reject any sample that is statistically dissimilar by applying a threshold. The so-called reject option was defined in [14] to indicate when the posterior probability of a sample x belonging to any class ω_i is such that $\forall i \in [1, 2, \cdots, K]$, $p(\omega_i|x) < T$ is less than a prescribed threshold T. This approach was found to be highly sensitive to the accuracy of the posterior estimates, and that a per-class threshold may improve performance [15], [16]. A direct approach to modeling the multivariate distribution of the known data is by constructing an empirical copula and then using this model to predict extreme events in subsequent data during operation or test to detect anomalies [17].

Although novelty detection is useful for accept/reject type of applications, it is not capable on its own in multi-category classification tasks, where the objective is to differentiate not only between known and unknown (anomalous) categories, but also to differentiate between different classes within the known categories [18]. Multi-category classifiers in supervised learning applications are commonly trained on data from a *closed set*. So, when these classifiers are presented with data that originates from a category outside of the training set, the classifier will potentially identify this sample as originating from one of the existing categories, and not uncommonly with high confidence [19], [20]. This can have devastating consequences, such as the time in 2016 when a lack of diverse training data resulted in a series of catastrophic failures in a CNN-based vehicular autopilot [21]. In contrast to closed set classification and reject/accept type anomaly detectors, *open set* recognition is the process of predicting both classes from the closed set and identifying anomalies originating from the space of unknown categories at query time. In [22], open set risk minimization was formulated as

$$\arg\min_{f\in\mathcal{H}} \left\{R_O(f) + \lambda R_\epsilon(f)\right\}, \tag{1}$$

which is a trade between empirical risk $R_\epsilon(f)$, or the risk of misclassification within the closed set, and open space risk $R_O(f)$, or the risk of assigning a label to the unknown space, with $\mathcal{H}$ representing the set of recognition functions. One approach to address the open set problem is the extreme value machine (EVM) which was derived using aspects of Extreme Value Theory (EVT) [23], [24], which has been shown to provide an abating bound on open set risk [25]. A core outcome of EVT is that the extreme values (tails) of a well-behaved continuous distribution can only assume a limited number of parametric forms, in particular, the Gumbel, Frechet and reverse Weibull distributions [26]. The EVM algorithm identifies the minimum (or transformed maximum) pairwise distance of a query point to the closest sample in the closed set and uses EVT to show that this distance follows a Weibull distribution. This enables the construction of an inclusion function to determine if the sample belongs to the class with the smallest pairwise distance, or belongs to the unknown (anomalous) class. The parameterization of the Weibull distribution and the choice of a statistical threshold δ is obtained using the closed set samples by parametric fitting and cross validation, respectively [27]. EVM represents one of the highest performing statistical approaches to open set recognition that is capable of kernel-free nonlinear variable bandwidth recognition in open set multi-category classification applications [28]. Like any of the approaches that model a distribution, there needs to be a sufficient and representative set of data for parametric fitting using EVT [29].

EVM, OC-SVM, IF, copula outlier detection (COPOD) and AE represent discriminative approaches to either open set recognition or accept/reject detectors. In this paper, we present a generative approach to open set recognition. The approach we take - distance-based probabilistic anomaly augemtation (DPAA) - directly addresses the open set recognition problem by reformulating (1) as a constrained optimization. The objective of DPAA is to minimize open space risk subject to an empirical risk constraint and can work with any classification algorithm. We comparatively demonstrate the open set recognition efficacy of DPAA using well-known multi-category data sets against state of the art discriminative and statistical approaches.

The rest of this paper is organized as follows. In Section II we review popular approaches to open set recognition. In Section III, we describe the DPAA algorithm and its formulation to address open set risk. In Section IV, we present a graphical assessment of how DPAA works, as wells as quantitative assessment of its performance on both open set recognition tasks as compared to EVM using an F_1-measure, as well as in accept/reject applications against OC-SVM, IF, COPOD and AE. We then conclude with a brief summary.

II. RELATED WORK

The term open set recognition was popularly coined by Schreirer in [22] to refer to the scenario in machine vision applications that not all classes present at query time are available during training. There has been a significant research thrust in the deep learning community to address this problem, and one early popular discriminative approach was OpenMax [30]. OpenMax calculates a 'mean activation' value derived from the penultimate layer of a deep learning network prior to the SoftMax output to generate an EVT-based weighting function. This weighting function modulates the SoftMax decision so that if an anomalous pattern were present, then the maximum categorical 'probability' would ideally be smaller than a threshold chosen to balance correct classification and open set rejection. An extension of OpenMax is Classification-Reconstruction learning for Open Set Recognition (CROSR) [31]. In CROSR, a deep hierarchical reconstruction net is formed in which intermediate layers of the network are compressed into latent space and reconstructed. CROSR uses this to construct a per class distance measure that is the L2 norm difference between both the activation and latent vectors and the per class means, using an OpenMax EVT-based framework to reject outliers. In the so-called Objectosphere approach [32], there are two modifications made to the loss function during training. The first is to define an entropic loss which treats known and unknown classes during training separately. Although the known class losses remain unchanged, unknown classes have their score uniformly distributed across all the known classes, i.e., the maximum entropy response. The second modification is to create an Objectosphere loss in which known classes that have small feature vector magnitudes that are inside the Objectosphere boundary are penalized, as are unknown classes with large ones. OpenMax, CROSR and Objectosphere each represent discriminative approaches to open set recognition in deep learning networks. Both OpenMax and CROSR leverages EVT, while Objectosphere modifies the loss function and requires that there be negative examples (outliers) during training.

A generative approach that builds on OpenMax is Generative OpenMax (G-OpenMax) for Multi-Class Open Set Classification [33]. Like OpenMax, G-OpenMax uses Weibull calibrated scores based on distances from the mean activation vectors in the penultimate layer of the network. However, in addition, G-OpenMax uses a conditional generative adversarial network (GAN) to create samples from the unknown category to generate well-calibrated probability scores for anomalies. Open generative adversarial networks (OpenGAN) [34] augments a classifier that already has access to open set samples with GAN generated data. Training is conducted in a manner that is similar to traditional GAN training. In the class conditioned auto-encoder (C2AE) [35], an encoder-classifier is trained in tandem and the weights are frozen. Then, the encoder (with weights frozen) and decoder are trained to generate images using a class conditional label that results in a large reconstruction error when the label does not match the class identity, and a small reconstruction error when it does. EVT is used to model the reconstruction errors with an associated threshold to identify outliers.

Each of the approaches to open set recognition described above are designed to work with deep learning networks that are principally operating on images. These discriminative and generative techniques, however, were not designed to work with other high-performance machine learning algorithms such as the Light Gradient Boosting Machine (LGBM) [36] in DSS applications. Further, generative models that rely on GANs are subject to instability during training, which may require access to negative examples (outliers) during training for stabilization [34]. There are also traditional approaches to open set recognition that rely on distance-based discrimination. The Nearest Non Outlier (NNO) [37] builds on the Nearest Class Mean (NCM) [38] classifier to identify both categorical samples and outliers based on their Euclidean distance. NNO uses the concept that non-negative combinations of abating functions (e.g., distances) can be thresholded to minimize open space risk. In [25], thresholded scores from a Weibull-calibratd SVM (W-SVM) are used to reject outliers. Traditional methods, however, may not represent the highest performing closed set classifier for the intended application.

TABLE 1. Symbolic notation definitions.

Symbol	Definition
$X = [x_1, \cdots, x_N]$	The closed set training samples, where $x_i = [x_{i,1}, \cdots, x_{i,D}]^T \in \mathbb{R}^D$ is a point in D-dimensional feature space
$\mathcal{X} = [\chi_1, \cdots, \chi_M]$	The test set consisting of the closed set of samples held out from the training set for classification performance evaluation
$C_{\mathcal{X}} = [c_1, \cdots, c_M]$	The categories (ground truth) associated with each sample in the test set
$\mathcal{A}_c = [a_1, \cdots, a_C]$	The set of synthetically generated anomaly candidates, where $a_i \in \mathbb{R}^D$
$\mathcal{A} = [a_1, \cdots, a_S]$	The set of accepted synthetically anomalies, where $S \leq C$ are the number of samples accepted
$D_{kNN}(x_i, x_{\forall j \neq i})$	The $k-$Nearest Neighbor (Euclidean) distance of x_i w.r.t. all *other* members of the training set
$D_{kNN}(a_i, X)$	The $k-$Nearest Neighbor (Euclidean) distance of a_i w.r.t. all members of the training set
$p(d \leq D_{kNN}(\cdot))$	The empirical cumulative distribution function (ECDF) of kNN distances
Q_1, Q_k, $Q_{1_{\max}}$, $Q_{1_{\min}}$, $Q_{k_{\max}}$, $Q_{k_{\min}}$	The quantile associated with kNN distances such that $\{d : p\left(d \leq D_{kNN}(x_i, x_{\forall j \neq i})\right) \geq Q_k\}$, Q_1 is with $k = 1$, and the min and max suffixes represent the maximum and minimum allowed values
$\alpha_1, \alpha_{\max}$	Scale factor to allow for 1NN distances between synthetic samples and training data that are greater than the maximum 1NN intra-training dataset distance
ϵ, Δ_ϵ	The decrease ϵ in classification accuracy after synthetic sample generation. The quantity Δ_ϵ helps to define a minimum classification error rate - $\epsilon - \Delta_\epsilon$ - after synthetic anomalies are included in the training set

III. TECHNICAL APPROACH

The DPAA algorithm generates synthetic anomalies at a statistically prescribed distance to the closed set boundary. This distance is balanced against a constraint that the empirical risk be no greater than a differential error rate ϵ, that is, the difference between classification performance on the closed set trained with and without anomalies. The distance used to generate and accept synthetic samples that represent anomalies is directly related to a divergence measure that relates the relative difference in the the anomalous and closed set distributions. DPAA directly addresses (1) by bringing samples in as close as possible to the boundary (irrespective of the boundary shape) to minimize open space risk, while ensuring that the differential error rate - or empirical risk - is no greater than ϵ.

To generate *candidate* synthetic anomalies, each of the data points in the closed set is used as a pivot, from which samples are randomly generated. The distance to the pivot from which samples are generated is directly related to the empirical distribution of distances within the closed set itself. The statistical measure of distances is used to quantify the (dis)similarity of the distribution of synthetic anomalies and closed set samples, and the choice of (dis)similarity is used in the process of minimizing the open set risk in (1). Because points in the 'interior' that are used as pivots will result in samples being generated inside the closed set as opposed to on or outside its boundary, sample generation is a two-step process: 1) generate candidate synthetic anomalies and 2) accept only those candidates on the boundary. The following subsections describe both the mechanics and mathematics of each of the points above in detail, and the notation used throughout is defined in Table 1.

A. CANDIDATE SYNTHETIC ANOMALY GENERATION

To generate synthetic anomalies, we use a 1-nearest neighbor (1NN) distance measure of an anomaly with respect to the closed set and the distribution of 1NN distances within the closed set itself. To that end, consider the empirical cumulative distribution (ECDF) of 1NN distances calculated from every training sample in the closed set to all other samples, i.e., $\{D_{1NN}(x_i, x_{\forall j \neq i})\}$, $\forall i \in [0, 1, \cdots, N]$. Using these distances, a single distance is selected which satisfies

$$d_p = \alpha_1 \cdot \min\left\{d : p(d \leq D_{1NN}(x_i, x_{\forall j \neq i})) \geq Q_1\right\}, \qquad (2)$$

where Q_1 and $\alpha_1 > 1$ are hyperparameters that are used to define the $1NN$ ECDF quantile and distance from that quantile, respectively. In (2), d_p is used as pivot distance from which synthetic anomalies are generated as

$$a_c = x_i + d_p \frac{\rho_c}{\|\rho_c\|^2} \in \mathcal{A}_c, \qquad (3)$$

where

$$\rho_c \sim \mathcal{N}(0, \mathrm{I}_D) \qquad (4)$$

is sampled from a D-dimensional isotropic Gaussian distribution. The intuition behind Eqs. (2), (3) and (4) is that a random D-dimensional sample is generated at a fixed pivot distance d_p with respect to each of the samples in the closed set. This distance is greater than or equal to the distance of the Q_1th quantile (depending on the value of α_1) of the closed set pivot and is instantiated at an angle $[\theta_1, \theta_2, \cdots, \theta_{D-1}]$ that is uniformly distributed over $(0, 2\pi]$ [39]. This process of generating anomaly candidates a_c is graphically illustrated in Fig. 1 for three cases of interest. The first case (C_1) has a synthetic sample generated at a distance d_p from the pivot x_i that lands squarely within the closed set boundary. In cases C_2 and C_3, the synthetic samples are are also generated at a distance d_p from the pivot, but in this case fall outside the closed set boundary with varying degrees of proximity. These three cases graphically illustrate the sample generation process to help visualize the intuition behind the sample anomaly generation and acceptance process, which is quantified in the next section. It is clear that the sample associated with C_1 landing within the boundary of the closed set might be mistaken for a closed set sample, or vice versa, increasing empirical risk. The anomalies C_2 and C_3 land at various distances outside the boundary, but are potentially too far away from the closed set boundary to mitigate open space risk. The question becomes which synthetically generated anomalies should be accepted and which should be rejected?

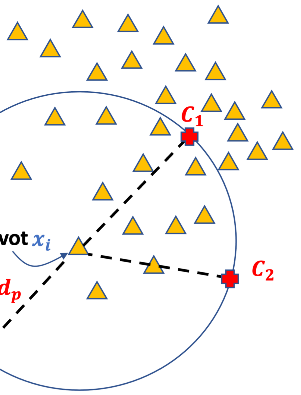

FIGURE 1. Various outcomes from synthetic anomaly sample generation. The triangles represent samples from the closed set, while crosses are examples of synthetically generated anomalies each landing at different angles at a distance of d_p from the pivot.

B. CANDIDATE SYNTHETIC ANOMALY ACCEPTANCE

The mechanics of candidate acceptance partially mirrors that of candidate generation, i.e., we look to find candidates whose *kNN* distance to the training set is *sufficiently* greater than the *kNN* distances within the training set. The mathematical justification for this observation [40] follows from divergence between the distribution of closed set sample $f(X)$ and the distribution of synthetic anomalies $f(\mathcal{A})$

$$D_{KL}(f(x)||f(\mathcal{A})) \sim \frac{D}{N}\sum_{i=1}^{S} \log \frac{D_{kNN}(x_i, \mathcal{A})}{D_{kNN}(x_i, x_{\forall j \neq i})} + K \qquad (5)$$

where the constant $K = N/(S-1)$. Eq. (5), which is an application of the general result in [41], states that a consistent estimate of the divergence between the training set multivariate distribution and the distribution of synthetically generated anomalies is directly related to *kNN* distances. Maximizing the numerator in (5), however, would push anomalies far away from the boundary of the training set, geometrically leaving a gap in open space into which anomalies could fall and remain undetected. To address this concern we consider the formulation of a constrained optimization to geometrically surround the (possibly non-convex) training set with anomalies without suffering a significant loss in core classification performance on the test set after training. To this end, and in an analogous manner to (2), we first define a decision rule with

$$d_k = \min\left\{d : p\left(d \leq D_{kNN}(x_i, x_{\forall j \neq i})\right) \geq Q_k\right\}, \qquad (6)$$

which represents the minimum distance from the set of *kNN* distances greater than or equal to the Q_k^{th} quantile, where synthetically generated candidates in $\mathcal{A}_c$ are accepted according to the rule:

$$\mathcal{A} = \{a_n : D_{kNN}(a_n, X) > d_k\}. \tag{7}$$

The problem then becomes one of deciding how to select an *optimal* α_1, Q_1 and d_p in (2) to generate synthetic candidates in (3), and d_k and Q_k to screen anomalous candidates in (7). To this end, let $f_{X\cup\mathcal{A}}(\cdot)$ and $f_X(\cdot)$ represent the classifiers trained with and without the synthetically generated anomalies, respectively, and $\mathbb{I}(\cdot)$ the indicator function defined as

$$\mathbb{I}(x) = \begin{cases} 1, & \text{if } x = 0 \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

with

$$P_X(\mathcal{X}) = \frac{\sum_{i=1}^{M} \mathbb{I}\left(f_X(\chi_i) - c_i\right)}{|\mathcal{C}_{\mathcal{X}}|}$$
$$P_{X\cup\mathcal{A}}(\mathcal{X}) = \frac{\sum_{i=1}^{M} \mathbb{I}\left(f_{X\cup\mathcal{A}}(\chi_i) - c_i\right)}{|\mathcal{C}_{\mathcal{X}}|} \tag{9}$$

where $|\cdot|$ represents that cardinality of the set with $P_X(\mathcal{X})$ and $P_{X\cup\mathcal{A}}(\mathcal{X})$ representing the classification accuracy on the test set. Then, with $c = [1, 1, 1]^T$, and $\theta = [\alpha_1, Q_1, Q_k]^T$, the constrained optimization

$$\begin{aligned} &\max_{\theta} c^T\theta \\ &\text{s.t. } P_{X\cup\mathcal{A}(\theta)}(\mathcal{X}) \geq P_X(\mathcal{X}) - \epsilon, \\ &\qquad 0 \leq Q_1, Q_k \leq 1, \quad \alpha_1 \in \mathbb{R}^+ \end{aligned} \tag{10}$$

finds the tightest boundary around the training set to place anomalies while ensuring that the classification performance on the test set when the classifier is trained with and without anomalies suffers a differential error rate no greater than ϵ. This effectively trades empirical risk (differential error rate) against open space risk (tightness of boundary) from (1) in the form of a constrained optimization in (10).

C. A PRACTICAL PROCEDURE FOR CANDIDATE GENERATION AND ACCEPTANCE

Although (10) quantifies a mathematically precise way of optimizing the generation of synthetic anomalies, a closed-form solution may not be possible due the fact that $P_{X\cup\mathcal{A}(\theta)}(\mathcal{X})$, which is calculated using a classifier trained using synthetically generated anomalies, is a highly nonlinear function of θ, cf. (2), (6), (7) and (9). To address this challenge, a fixed point iterative approach to approximately solve (10) is summarized in Algorithm 1. The parameters that control the generation and placement of synthetic anomalies are the same as those in (10), namely α_1, Q_1 and Q_k. These parameters form the core search space over which Alg. 1 operates. The parameters α_1 and Q_1 influence the generation of samples in (2), where increasing these parameters push samples in $\mathcal{A}_c$ further from the boundary of the training set, making it more likely the candidates are accepted into the set $\mathcal{A}$. In concert with α_1 and Q_1, Q_k controls the likelihood of candidates from $\mathcal{A}_c$ being accepted into $\mathcal{A}$. But, unlike α_1 and Q_1, decreasing Q_k increases the likelihood of candidate selection.

Algorithm 1 Generate Synthetic Anomaly Sample Set $\mathcal{A}$

```
Require:
  $k, \epsilon, \Delta_\epsilon, Q_1, Q_k, Q_{1_{\max}}, Q_{k_{\max}}, Q_{1_{\min}}, Q_{k_{\min}}, \alpha_1, \alpha_{1_{\max}}$
  while $Q_{k_{\min}} \leq Q_k \leq Q_{k_{\max}}$ do
    In (6) compute $d_k$
    while $\alpha_1 \leq \alpha_{1_{\max}}$ do
      In (2) compute $d_p$
      From (3) compute candidate anomaly samples $\mathcal{A}_c$
      Accept candidate to form $\mathcal{A}$ using (7) ((12) for $A_K$)
      if $|\mathcal{A}| >$ minCandSamples then
        break
      else if $Q_1 \leq Q_{1_{\max}}$ then
        $Q_1 \leftarrow (Q_1 + Q_{1_{\max}})/2$
      else
        $\alpha_1 \leftarrow (\alpha_1 + \alpha_{1_{\max}})/2$
      end if
    end while
    if $\epsilon - \Delta_\epsilon < P_{X\cup\mathcal{A}}(\mathcal{X}) - P_X(\mathcal{X}) < \epsilon$ then
      break
    else if $P_{X\cup\mathcal{A}}(\mathcal{X}) - P_X(\mathcal{X}) \leq \epsilon - \Delta_\epsilon$ then
      $Q_{k_{\min}} \leftarrow Q_k$
      $Q_k \leftarrow (Q_k + Q_{k_{\max}})/2$
    else
      $Q_{k_{\max}} \leftarrow Q_k$
      $Q_k \leftarrow (Q_k + Q_{k_{\min}})/2$
    end if
  end while
```

Consider the first the adaptive selection of Q_1 and α_1. A modified binary search for the values Q_1 and α_1 is used to find the minimum number of samples that meet the candidate selection requirement as specified in (7). Once the candidates are selected, a test is used to determine if the differential error is within the range $[\epsilon - \Delta_\epsilon, \epsilon]$. If the differential error is outside of this range, then Q_k is adjusted. This is in contrast to (10), where only a single differential error is specified. The rationale is as follows. In the constrained optimization approach as specified in (10), minimizing Q_1, α_1 and Q_k brings samples in towards the boundary, and this is balanced against the constraint that the addition of synthetic samples in the training dataset preserve core classification performance, i.e., the differential error is no greater than ϵ. This is approximately captured in the upper and lower bound on the differential error in Alg. 1. If the upper bound ϵ is violated, Q_k is adaptively increased which forces the selection of samples that are further away from the boundary of the test set. Conversely, if the minimum error $\epsilon - \Delta_\epsilon$ is violated, then the samples are selected that lie closer to the boundary by reducing Q_k. A binary search is used to adaptively adjust Q_k so that samples fall within differential error window $[\epsilon - \Delta_\epsilon, \epsilon]$. As $\Delta_\epsilon \to 0$, Alg. 1 more closely approximates 10, but this comes at the expense of an increase in the search time. We have found that in practice a Δ_ϵ that is between 25% and 35% of ϵ works well in practice, and this is topic is expanded on further in Section IV where results are presented. The inialization parameters

D. CATEGORICAL DATA AND NORMALIZATION

Data normalization is commonly employed in machine learning when the features used for training and classification are on different scales. In the examples that follow, we use standard normalization with

$$x_i^{nrm} = \frac{x_i - \mu_i}{\sigma_i}, \quad \chi_i^{nrm} = \frac{\chi_i - \mu_i}{\sigma_i}, \tag{11}$$

where μ_i is the mean value of feature i and σ_i its standard deviation. Some data sets will also include categorical features, i.e., features that can only take on a finite and discrete set of values. In this case, the features generated using (3) are snapped back to a grid after acceptance using (7) such that

$$\mathcal{A}_K = \left\{ x_{i,j} : \|a_n - x_{i,j}\| < \|a_n - x_{i,k}\|, \forall j \neq k \right\}, \tag{12}$$

where $a_n \in \mathcal{A}_K \in \mathcal{A}$ are the subset of columns corresponding to range of possible choice for categorical, integer and discrete valued features.

E. DISCUSSION

The question of how to generate and where to place anomalies is not fully answered by (5), but this does offer quantitative evidence for the idea that - from an information-theoretic perspective - *kNN* distance is a measure of statistical (dis)similarity. Indeed, maximizing (5) will result in samples with maximum dissimilarity but this would leave a region in feature space where an anomaly encountered in the field would have a KL divergence measure 'closer' to the training set than that of the synthetically generated anomalies. Instead, we consider the question of how to minimize the numerator in (5) (given the denominator is fixed) while maximizing the likelihood that samples from the (closed) test set are correctly classified. In theory, it is possible to turn this into a constrained optimization, with the constraint that that differential error rate between a classifier trained with and without synthetic anomalies is less than a prescribed threshold ϵ, cf. (10). This quantitatively addresses the question posed in Sec. III-B of how to find candidates with *kNN* distances *sufficiently* different from those in the training set, while also balancing the empirical risk. The choice to use accuracy in (9) could be replaced by any other measure such an F_β-measure (e.g., F_1) or AUC. Given the highly nonlinear nature of the constrained optimization, the pseudo-code for a computationally efficient approximation to (10) is presented in Alg. 1, and this is the algorithm which was used to obtain the results presented in Section IV.

IV. COMPARATIVE PERFORMANCE

A. VISUALIZING ANOMALY GENERATION

To get a visual sense of how anomalies are generated and the impact that the choice of fixed and searchable hyperparameters have on performance, we've used the so-called Banana dataset from Google's standard classification library [42] which is plotted for reference in Fig. 2. There are two categories - 0 and 1 - in the Banana dataset, with each category containing roughly 2500 samples for a combined dataset size of 5000 samples. A total of 5000 synthetic anomalies were generates to match the size of the banana data set. To derive the baseline performance $P_X(\mathcal{X})$, an XGBoost classifier [43] was trained and tested on the two class data set, with a training set X representing 80% of samples chosen at random, and a test set $\mathcal{X}$ made up of the remaining 20% of the samples. The first set of synthetic samples $\mathcal{A}$ derived from the Banana data set is plotted in Fig. 3.

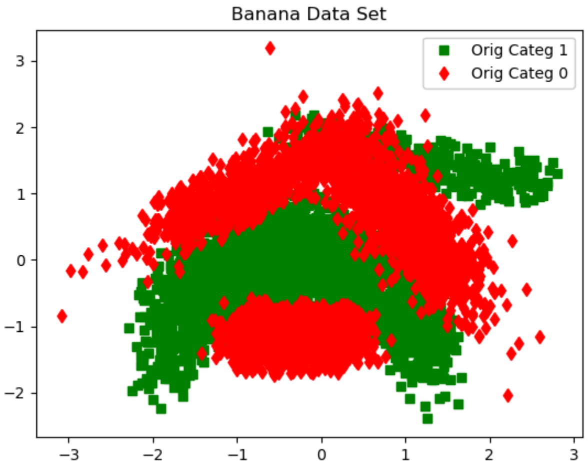

FIGURE 2. Banana dataset.

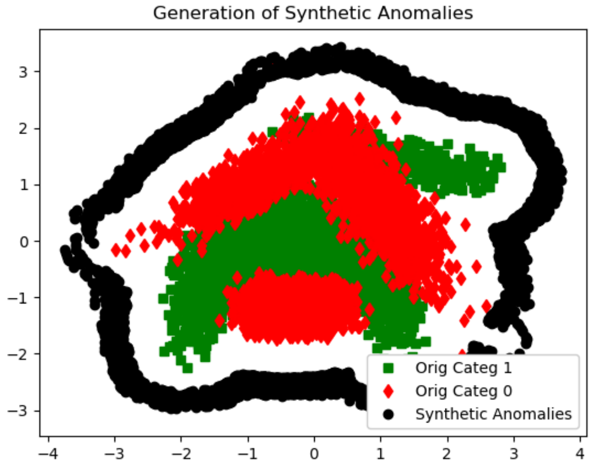

FIGURE 3. Banana dataset with synthetic anomalies generated using $\epsilon = .01$, $\Delta_\epsilon = .0035$, $\alpha_1 = 1$, $Q_1 = 1$, $Q_k = .9995$.

The differential error $\epsilon = .01$ and offset $\Delta_\epsilon = .0035$ were chosen resulting in a differential error range of [.0065, .01%], and the directed optimization of Alg. 1 recovered the ECDF quantile parameters $Q_1 = 1$, $Q_k = 0.9995$ for $k = 4$ and an $\alpha_1 = 1$. The synthetic anomalies generated all landed on the boundary of the Banana data set, but at a sufficient distance so that classifier (in this case XGBoost) could differentiate them from the the original 2-categories $\{0, 1\}$ at a loss no greater than $\epsilon = .01$ in classification performance. The second set

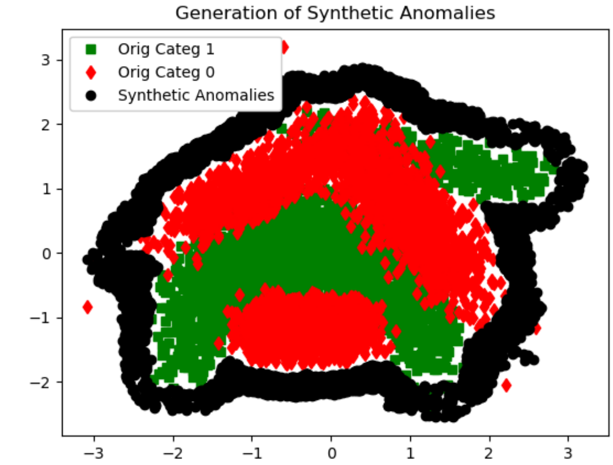

FIGURE 4. **Banana dataset with synthetic anomalies generated using $\epsilon = .05$, $\Delta_\epsilon = .0175$, $\alpha_1 = 1$, $Q_1 = 1$, $Q_k = .995$.**

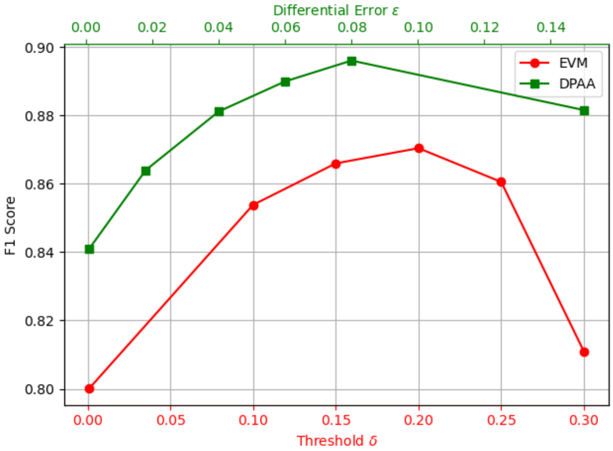

FIGURE 5. **Relative performance of open set recognition for both DPAA and EVM on the OLETTER data set. F_1 performance for DPAA is measured with respect to the top axis in terms of a differential error rate. For EVM, F_1 performance is measured with respect to the bottom axis in terms of a probability threshold δ as described in [24].**

TABLE 2. **DPAA initialization parameters to obtain the F_1-scores and P_d vs. P_{FA} in Figs. 5 through 10.**

Parameter	Value
Q_1	1.0
$Q_{1_{\min}}$	0.4
$Q_{1_{\max}}$	1.0
Q_k	0.8
$Q_{k_{\min}}$	0.2
$Q_{k_{\max}}$	1.0
α_1	1
$\alpha_{\max}$	1.5
k	4

of synthetic samples derived from the Banana data set is plotted in Fig. 4, but this time with a differential error range of [.0325, .05], which resulted in the ECDF quantile parameters $Q_1 = 1$, $Q_k = 0.995$ and an $\alpha_1 = 1$. In this case, synthetic anomalies tightly hugged the boundary of the Banana data set, resulting in more samples from the test set $\mathcal{X}$ being mistaken for anomalies, but at a rate no greater than $\epsilon = 0.05$.

B. QUANTITATIVE ASSESSMENT - OPEN SET RECOGNITION

Although the banana data set was useful to help visualize how the DPAA algorithm works, it is not sufficient to measure DPAA performance. Therefore, to test the efficacy of DPAA, we compared its performance to that of EVM operating on high dimensional mutli-category datasets using an F_1-measure as a function of precision and recall, which is defined as

$$Precision = \frac{TP}{TP + FP},$$
$$Recall = \frac{TP}{TP + FN}, \tag{13}$$

where TP, FP, FN represent the true positive, false positive to false negative rates, respectively, and from which the harmonic mean - or F_1 score

$$F_1 = \frac{2}{1/Precision + 1/Recall}, \tag{14}$$

is derived. In addition to the F_1-score, we also measure the ability of DPAA to both detect anomalies (probability of detection, or P_D) and to not falsely assign samples as anomalies (probability of false alarm, or P_{FA}). We consider P_D and P_{FA} separately from the overall F_1-measure as a function of precision and recall for two reasons. First, there may be a significant difference in *cost* associated with how anomalies are processed, both in terms of their detection and misclassification. Second, P_D and P_{FA} are reflective of an accept/reject type of detector as discussed in Section I, and not multi-category classification which includes an outlier category. To start, we used the OLETTER dataset developed in [25] which was used to demonstrate EVM's performance in open set recognition in [24]. The test we performed started with training both EVM and DPAA on 20- of the 26-letters selected at random, where the 6-letters that were held-out from the training set were included in the testing set. An 80/20 train/test split was used for the 20 letters prior to the addition of the 6-letters in the test set, and 10-fold cross validation was used to obtain average performance results. The EVM code used for both hyper-parameter optimization and to obtain the results that follow were obtained from the authors' github repository [44]. Like for the Banana dataset, DPAA used XGBoost for $f_X(\cdot)$ and $f_{X\cup\mathcal{A}}(\cdot)$ in (9) to measure differential error, $P_X(\mathcal{X}) - P_{X\cup\mathcal{A}(\theta)}(\mathcal{X})$. The open set performance of DPAA and EVM operating on the OLETTER test set after training is plotted in Fig. 5.

To obtain results for DPAA, we initialized the parameters in Alg. 1 with the those listed in Table 2 and varied the differential error with values in Table 3 to obtain the results in Figs. 5 - 10. For comparison, we varied the EVM threshold

TABLE 3. DPAA differential error (ϵ) and EVM threshold parameters (δ) used to obtain the F_1-scores in Figs. 5 through 10.

DPAA	EVM
$[\epsilon - \Delta_\epsilon, \quad \epsilon]$	δ
$[0, \quad 10^{-3}]$	10^{-3}
[0.013, 0.020]	0.100
[0.026, 0.040]	0.150
[0.039, 0.060]	0.200
[0.052, 0.800]	0.250
[0.098, 0.150]	0.300

δ in [24] using

$$\hat{c}_i = \begin{cases} \arg\max_{m} p(c_m|\chi_i), & \text{if } p(c_m|\chi_i) \geq \delta \\ \text{Anomaly}, & \text{Otherwise}, \end{cases} \tag{15}$$

which has the effect of trading off P_D for P_{FA} in an analogous manner to ϵ for DPAA. In general, both DPAA and EVM followed similar trajectories in terms of an F_1-score in the open set recognition task, but with DPAA outperforming it in all cases.

Although the F_1-score is a comprehensive snapshot of open set recognition, it was also of interest to measure the performance of both algorithms in an accept/reject framework. Because of the wide availability of curated outlier detection algorithms that come bundled in the PyPI 3.7 library PyOD [45], we chose to include a representative set of 4 of these PyOD routines for comparison: Isolation Forest (IF), AutoEncoder (AE), One Class Support Vector Machine (OC-SVM) and Copula Outlier Detection (COPOD). Each of these PyOD routines had a detection threshold that was controlled by a single parameter, *contamination*, which for our tests varies from 0 to 0.175 in steps of .025. The AE feedforward deep learning architecture was modified from the default to fit the dimensionality of the OLETTER dataset, such that the number of hidden neurons per layer were [15, 8, 4, 8, 15]. In all other cases, default parameter settings were used for IF, AE, OC-SVM and COPOD. The probability of detection versus false alarm for all routines is plotted in Fig. 6. It was interesting to note that the PyOD routines, whose only objective was accept/reject, performed quite poorly relative to both EVM and DPAA on this dataset. Similar to the open set recognition performance using an F_1 score, DPAA detection performance was greater than or equal to that of EVM for any chosen P_{FA} in our tests.

It is interesting to take a deeper look at open set recognition performance of DPAA. In Fig. 5, DPAA has its peak F_1 performance at nearly 90%, with a corresponding differential error rate of roughly 8%. One may ask, how is it that a relatively high differential error rate leads to the highest F_1-score? The reason is that the differential error rate measure is w.r.t. to only the closed set samples in the test set, while the F_1 measure is w.r.t. both the closed set samples and anomalies. This observation will hold in the results that follow.

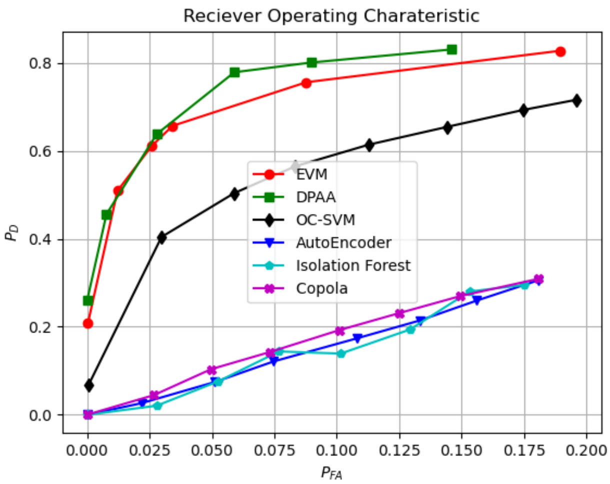

FIGURE 6. Receiver operating characteristic of an accept/reject configuration for anomalie on the OLETTER data set. Here both DPAA and EVM performance is compared to that of select outlier detection routines in the Python Outlier Detection (PyOD) library.

In addition to the OLETTER dataset, we tested DPAA on the so called multi-feature Fourier (mfeat-fourier) multi-category data set, which is publicly available on the UCI curated multi-category classification website [46]. The mfeat-fourier dataset has 76 features describing the handwritten digits 0-9, and serves as an analogue to the alphabetic OLETTER dataset. Similarly to OLETTER open set testing, we randomly removed a single digit from the training set and used it in testing. Like OLETTER, an 80/20 train/test split was used for the 9 handwritten digits prior to the addition of the 10th digit in the test set, and 10-fold cross validation was used to obtain average performance results. The open set performance of DPAA and EVM operating on the mfeat-fourier test set after training is plotted in Fig. 7. As was true for the OLETTER data set, DPAA outperformed EVM at nearly all cases. For an accept/reject mode of operation DPAA achieved a higher detection rate P_d for a given false alarm rate P_{FA} than both EVM and all of the routines from the PyOD library. In this case, however, OC-SVM performance was comparable to that of EVM and significantly outperformed the other outlier detection algorithms.

C. CATEGORICAL DATA

Finally, we consider the problem of intrusion detection from both an accept/reject and open set recognition perspective using the NSL-KDD dataset [47]. This data represents a distinctly different one from either the OLETTER or multi-feature Fourier character based datasets given all of the features are either categorical, discrete or integer in nature, and we leverage (12) in the process of anomaly generation. For this data set, in addition to 'Normal' network traffic, we chose to include Probe, Denial of Service (DoS) and Remote-to-Local types of attacks with attack types tabulated

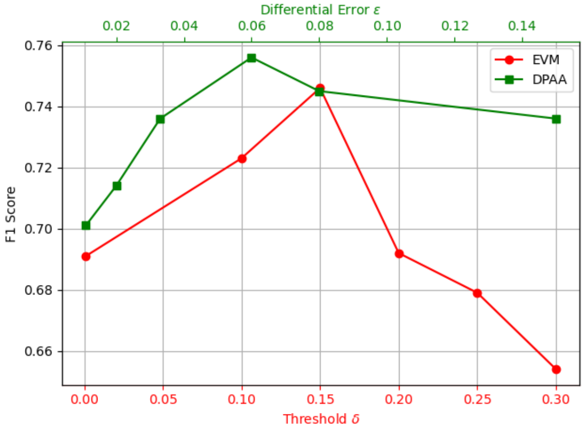

FIGURE 7. Relative performance of open set recognition for both DPAA and EVM on the multi-feature Fourier data set. F_1 performance for DPAA is measured with respect to the top axis in terms of a differential error rate. For EVM, F_1 performance is measured with respect to the bottom axis in terms of a probability threshold δ as described in [24].

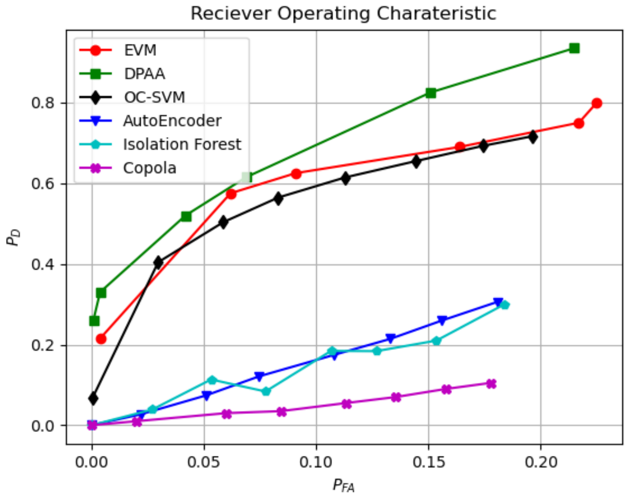

FIGURE 8. Receiver operating characteristic of an accept/reject configuration for anomalies on the multi-feature Fourier data set. Here both DPAA and EVM performance is compared to that of select outlier detection routines in the Python Outlier Detection (PyOD) library.

in Table 4 in the training set. The DoS attack *pod* was held out from the training set and included in the test set as it is one of the most challenging attacks to identify. We used an identical train/test split and cross validation procedure to the one used for both the OLETTER and multi-feature Fourier data sets, with the exception that *pod* was always held out during training but included during test. The F_1 performance of both DPAA and EVM are plotted in Fig. 9, with DPAA significantly outperforming EVM in recognizing both normal traffic as well as the 14 categories of attack including *pod*, which as previously mentioned was held from the training set but included in the test set.

We also ran both DPAA and EVM in an accept/reject mode of operation and compared their performance with respect to the four other PyOD routines with the results plotted in Fig. 10. Both DPAA and EVM had a significantly lower false alarm rate than OC-SVM, IF, AE, and COPOD making them both far more useful in scenarios where a relatively low false alarm is critical for successful system operations where a significant number of normal packets aren't inadvertently blocked.

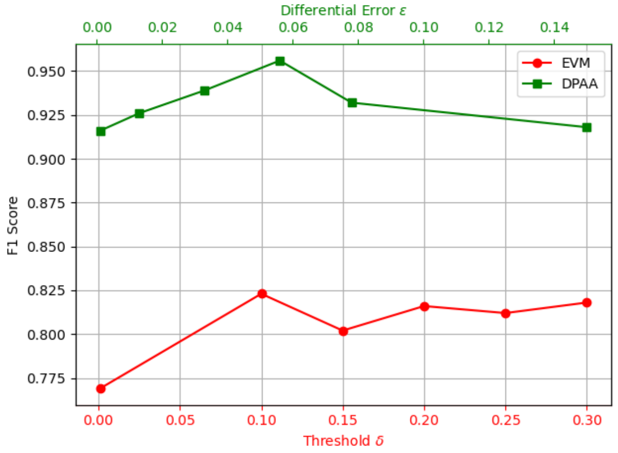

FIGURE 9. Relative performance of open set recognition for both DPAA and EVM on the NSL-KDD intrusion detection data set. F_1 performance for DPAA is measured with respect to the top axis in terms of a differential error rate. For EVM, F_1 performance is measured with respect to the bottom axis in terms of a probability threshold δ as described in [24].

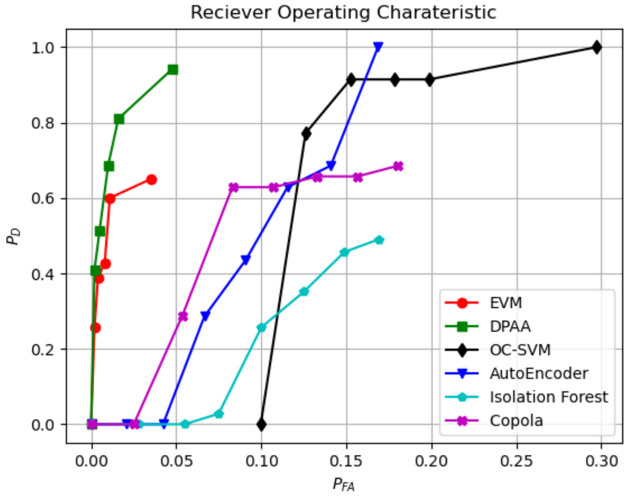

FIGURE 10. Receiver operating characteristic of an accept/reject configuration for anomalies on the NSL-KDD intrusion detection dataset. Here both DPAA and EVM performance is compared to that of select outlier detection routines in the Python Outlier Detection (PyOD) library.

TABLE 4. Types of attacks used during training. Note that the DoS attack, *pod*, was held out and used during test.

Probe	DoS	R2L
portsweep	neptune	guess passwd
ipsweep	smurf	warezclient
satan	apache2	warezmaster
nmap	back	
mscan	teardrop	

V. SUMMARY

This paper presented a distance-based probabilistic anomaly augmentation (DPAA) approach to address the open set recognition problem. DPAA generates samples to encapsulate the (possibly non-convex) training a set in feature space. This generative approach is directly formulated to optimize open space risk subject to an empirical risk constraint, whose optimization mechanics are grounded in an information theoretic and statistical measure of closeness. Using a representative ensemble of data sets, DPAA demonstrated superior performance both in novelty detection and open set recognition against some of the highest performing state-of-the-art algorithms with the flexibility to work in concert with any classifier.

APPENDIX

LIST OF ACRONYMS

AE	AutoEncoder
C2AE	Class Conditional AutoEncoder
COPOD	Copula Outlier Detector
CROSR	Classification and Reconstruction Open Set Recognition
DNN	Deep Neural Network
DPAA	Distance-based Probabilistic Anomaly Augmentation
ECDF	Empirical Cumulative Distribution Function
EVM	Extreme Value Machine
EVT	Extreme Value Theory
IF	Isolation Forest
KL	Kullback-Leibler
LGBM	Light Gradient Boosting Machine
NCM	Nearest Class Mean
NNO	Nearest Non Outlier
OC-SVM	One Class Support Vector Machine
PyOD	Python Outlier Detector
XGBoost	Extreme Gradient Boosting

ACKNOWLEDGMENT

Opinions, interpretations, recommendations, and conclusions are those of the authors are not necessarily endorsed by the U.S. Government.

REFERENCES

[1] L. Tarassenko, "Novelty detection for the identification of masses in mammograms," in *Proc. 4th Int. Conf. Artif. Neural Netw.*, 1995, pp. 442–447.

[2] J. Henrydoss, S. Cruz, E. M. Rudd, M. Gunther, and T. E. Boult, "Incremental open set intrusion recognition using extreme value machine," in *Proc. 16th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2017, pp. 1089–1093.

[3] P. K. Chan, W. Fan, A. L. Prodromidis, and S. J. Stolfo, "Distributed data mining in credit card fraud detection," *IEEE Intell. Syst. Appl.*, vol. 14, no. 6, pp. 67–74, Nov./Dec. 1999.

[4] S. Nandi, H. A. Toliyat, and X. Li, "Condition monitoring and fault diagnosis of electrical motors—A review," *IEEE Trans. Energy Convers.*, vol. 20, no. 4, pp. 719–729, Dec. 2005.

[5] C. Geng, S.-J. Huang, and S. Chen, "Recent advances in open set recognition: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3614–3631, Oct. 2021.

[6] M. A. F. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Process.*, vol. 99, pp. 215–249, Jun. 2014.

[7] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, Jul. 2014.

[8] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 413–422.

[9] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[10] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Jul. 1995.

[11] B. B. Thompson, R. J. Marks, J. J. Choi, M. A. El-Sharkawi, M.-Y. Huang, and C. Bunje, "Implicit learning in autoencoder novelty assessment," in *Proc. Int. Joint Conf. Neural Netw.*, vol. 3, May 2002, pp. 2878–2883.

[12] H. Xiang, J. Wang, K. Ramamohanarao, Z. Salcic, W. Dou, and X. Zhang, "Isolation forest based anomaly detection framework on non-IID data," *IEEE Intell. Syst.*, vol. 36, no. 3, pp. 31–40, May 2021.

[13] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, 2009.

[14] C. Chow, "On optimum recognition error and reject tradeoff," *IEEE Trans. Inf. Theory*, vol. IT-16, no. 1, pp. 41–46, Jan. 1970.

[15] G. Fumera, F. Roli, and G. Giacinto, "Reject option with multiple thresholds," *Pattern Recognit.*, vol. 33, no. 12, pp. 2099–2101, Dec. 2000.

[16] D. M. J. Tax and R. P. W. Duin, "Growing a multi-class classifier with a reject option," *Pattern Recognit. Lett.*, vol. 29, no. 10, pp. 1565–1570, Jul. 2008.

[17] Z. Li, Y. Zhao, N. Botta, C. Ionescu, and X. Hu, "COPOD: Copula-based outlier detection," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2020, pp. 1118–1123.

[18] Y. Fu, T. Xiang, Y.-G. Jiang, X. Xue, L. Sigal, and S. Gong, "Recent advances in zero-shot recognition: Toward data-efficient understanding of visual content," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 112–125, Jan. 2018.

[19] A. Khamis, Z. Ismail, K. Haron, and A. Mohammed, "The effects of outliers data on neural network performance," *J. Appl. Sci.*, vol. 5, no. 8, pp. 1394–1398, Jul. 2005.

[20] P. Stock and M. Cisse, "ConvNets and ImageNet beyond accuracy: Understanding mistakes and uncovering biases," in *Computer Vision*, vol. 11210. Cham, Switzerland: Springer, 2018, 2008.

[21] N. Sünderhauf, O. Brock, W. Scheirer, R. Hadsell, D. Fox, J. Leitner, B. Upcroft, P. Abbeel, W. Burgard, M. Milford, and P. Corke, "The limits and potentials of deep learning for robotics," *Int. J. Robot. Res.*, vol. 37, nos. 4–5, pp. 405–420, 2018.

[22] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1757–1772, Jul. 2013.

[23] S. Kotz and S. Nadarajah, *Extreme Value Distributions: Theory and Applications*. Singapore: World Scientific, 2000.

[24] E. M. Rudd, L. P. Jain, W. J. Scheirer, and T. E. Boult, "The extreme value machine," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 762–768, Mar. 2018.

[25] W. J. Scheirer, L. P. Jain, and T. E. Boult, "Probability models for open set recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2317–2324, Nov. 2014.

[26] L. de Haan and A. Ferreira, *Extreme value theory: An introduction*. Cham, Switzerland: Springer, 2007.

[27] W. J. Scheirer, A. Rocha, R. J. Micheals, and T. E. Boult, "Meta-recognition: The theory and practice of recognition score analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1689–1695, Aug. 2011.

[28] E. Vignotto and S. Engelke, "Extreme value theory for anomaly detection—The GPD classifier," *Extremes*, vol. 23, no. 4, pp. 501–520, Dec. 2020.

[29] M. Aitkin and D. Clayton, "The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM," *Appl. Statist.*, vol. 29, no. 2, p. 156, 1980.

[30] A. Bendale and T. E. Boult, "Towards open set deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1563–1572.

[31] R. Yoshihashi, W. Shao, R. Kawakami, S. You, M. Iida, and T. Naemura, "Classification-reconstruction learning for open-set recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4011–4020, doi: 10.1109/cvpr.2019.00414.

[32] A. R. Dhamija, M. Günther, and T. E. Boult, "Reducing network agnostophobia," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2018, pp. 9175–9186.

[33] Z. Ge, S. Demyanov, Z. Chen, and R. Garnavi, "Generative OpenMax for multi-class open set classification," 2017, *arXiv:1707.07418*.

[34] S. Kong and D. Ramanan, "OpenGAN: Open-set recognition via open data generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 793–802.

[35] P. Oza and V. M. Patel, "C2AE: Class conditioned auto-encoder for open-set recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2302–2311.

[36] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 3149–3157.

[37] A. Bendale and T. Boult, "Towards open world recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1893–1902.

[38] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka, "Distance-based image classification: Generalizing to new classes at near-zero cost," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2624–2637, Nov. 2013.

[39] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*. New York, NY, USA: Wiley, 2005.

[40] J. Goodman, S. Sarkani, and T. Mazzuchi, "Distance-based probabilistic data augmentation for synthetic minority oversampling," *ACM/IMS Trans. Data Sci.*, Feb. 2022, doi: 10.1145/3510834.

[41] Q. Wang, S. R. Kulkarni, and S. Verdu, "Divergence estimation for multidimensional densities via k-nearest-neighbor distances," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2392–2405, Apr. 2009.

[42] S. Jaichandaran. (2020). *Standard Classification Library Banana Data Set*. [Online]. Available: https://www.kaggle.com/saranchandar/standard-classification-with-banana%-dataset

[43] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[44] EMRResearch. *Extreme Value Machine*. Accessed: Nov. 24, 2021. [Online]. Available: https://github.com/EMRResearch/ExtremeValueMachine

[45] Y. Zhao, Z. Nasrullah, and Z. Li, "PyOD: A Python toolbox for scalable outlier detection," *J. Mach. Learn. Res.*, vol. 20, no. 96, pp. 1–7, Jan. 2019. [Online]. Available: http://jmlr.org/papers/v20/19-011.html

[46] D. Dua and C. Graff. (2017). *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml

[47] G. Mohi-Ud din. (2018). *NSL-KDD*. [Online]. Available: https://dx.doi.org/10.21227/425a-3e55

JOEL GOODMAN (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Boston University. He is currently pursuing the Ph.D. degree in systems engineering with The George Washington University, Washington, DC, USA.

He was previously employed with Eastman Kodak, worked for venture capital-funded Hammerhead Networks (later acquired with Cisco Systems), and was a Technical Staff Member at the MIT Lincoln Laboratory (MIT LL), most recently serving on the Chief Technology Officer's Technical Advisory Group. He is also a Senior Research Engineer with the Tactical Electronic Warfare Division, U.S. Naval Research Laboratory.

Mr. Goodman was a member of the Organizing Committee of the IEEE GlobalSIP 2016. He was a recipient of the Eastman Technical Achievement Award for his work on magnetic imaging, the 2008 MIT LL Team Excellence Award for his work on nonlinear equalization, the 2012 NRL Alan Berman Research Publication Award, and the NRL Review Award. He is also serving as the Chair for the IEEE Computational Intelligence Society Chapter, Washington. He worked as a Lecturer at the 2014 Virginia Tech Symposium and Summer School on Wireless Communications. He was an Invited Lecturer with the IEEE Advanced Signal Processing Symposium on the topic of nonlinear signal processing, in 2008.

SHAHRAM SARKANI received the B.S. and M.S. degrees in civil engineering from Louisiana State University, Baton Rouge, LA, USA, and the Ph.D. degree in civil engineering from Rice University, Houston, TX, USA. He is currently a Professor of engineering management and systems engineering with The George Washington University, Washington, DC, USA. His current administrative appointments are inaugural Director of the School of Engineering and Applied Science Off-Campus and Professional Programs (since 2016), the school unit to establish cross-disciplinary and departmental programs for offer off-campus and/or by synchronous distance learning; and a Faculty Adviser and an Academic Director of the EMSE Off-Campus Programs (since 2001), the department unit that designs and administers five separate graduate degree programs in six areas of study that enroll more than 800 students across the USA and abroad. He joined GW, in 1986, where previous administrative appointments include the Chair of the Civil, Mechanical, and Environmental Engineering Department (1994–1997); and an Interim Associate Dean for Research of the School of Engineering and Applied Science (1997–2001). In more than 500 technical publications and presentations, his research in systems engineering, systems analysis, and applied enterprise systems engineering has application to risk analysis, structural safety, and reliability. He has conducted sponsored research for such organizations as NASA, NIST, NSF, U.S. AID, and the U.S. Departments of Interior, Navy, and Transportation. He was inducted into the Civil and Environmental Engineering Hall of Distinction, Louisiana State University, in 2010; and was awarded the Walter L. Huber Civil Engineering Research Prize by the American Society of Civil Engineers in 1999. He is a Registered Professional Engineer in Virginia.

THOMAS MAZZUCHI received the B.A. degree in mathematics from the Gettysburg College, Gettysburg, PA, USA, in 1978, and the M.S. and D.Sc. degrees in operations research from The George Washington (GW) University, Washington, DC, USA, in 1979 and 1982, respectively. He is currently a Professor of engineering management and systems engineering, and the Chair of the Department of Engineering Management and Systems Engineering, School of Engineering and Applied Science, GW. Formerly, he was the Chair of the Department of Operations Research, and as Interim Dean of the School of Engineering and Applied Science. He has been engaged in consulting and research in the areas of reliability and risk analysis, and systems engineering techniques, for more than 30 years. He served for two and a half years as a Research Mathematician with the International Operations and Process Research Laboratory, Royal Dutch Shell Company. While at Shell, he was engaged in reliability and risk analysis of large processing systems, maintenance optimization of off-shore platforms, and quality control procedures at large-scale chemical plants. In his academic career, he has held research contracts in development of testing procedures for both the U.S. Air Force and the U.S. Army; in spares provisioning modeling with the U.S. Postal Service; in mission assurance with NASA; and in maritime safety and risk assessment with the Port Authority of New Orleans, the Washington Office of Marine Safety, the Washington State Department of Transportation, and the San Francisco Bay Area Transit Authority. He is an Elected Member of the International Statistics Institute.