Detecting Irregular Network Activity with Adversarial Learning and Expert Feedback

Gopikrishna Rathinavel
Virginia Tech
Blacksburg, VA
rgopikrishna@vt.edu

Nikhil Muralidhar
Stevens Institute of Technology
Hoboken, NJ
nmurali1@stevens.edu

Timothy O'Shea
DeepSig Inc & Virginia Tech
Arlington, VA
tim@deepsig.io

Naren Ramakrishnan
Virginia Tech
Arlington, VA
naren@cs.vt.edu

***Abstract*—Anomaly detection is a ubiquitous and challenging task, relevant across many disciplines. With the vital role communication networks play in our daily lives, the security of these networks is imperative for the smooth functioning of society. To this end, we propose a novel self-supervised deep learning framework CAAD for anomaly detection in wireless communication systems. Specifically, CAAD employs contrastive learning in an adversarial setup to learn effective representations of normal and anomalous behavior in wireless networks. We conduct rigorous performance comparisons of CAAD with several state-of-the-art anomaly detection techniques and verify that CAAD yields a mean performance improvement of 92.84%. Additionally, to adapt to the dynamic shifts in benign and anomalous data distributions, we also augment CAAD enabling it to systematically incorporate expert feedback through a novel contrastive learning feedback loop to improve the learned representations and thereby reduce prediction uncertainty (CAAD-EF). We view CAAD-EF as a novel, holistic, and widely applicable solution to anomaly detection. Our source code and data are available online[1]**

***Index Terms*—Anomaly detection, Generative Adversarial Networks, Wireless, Self-supervised learning, Contrastive Learning**

I. Introduction

Wireless communications systems form an essential component of cyber-physical systems in urban environments. They enable us to access the internet and connect with others remotely, thereby serving as a vital means of human interaction. They also connect thousands of sensors, applications, industrial networks, critical communications systems, and other infrastructure. Hence, state monitoring and detection of irregular activity in wireless networks are essential to ensuring robust and resilient system operational capabilities.

The *electromagnetic spectrum* (simply referred to as 'the spectrum') is the information highway through which most forms of electronic communication occur. Parts of the spectrum are grouped into 'bands' (based on the wavelength) which can be thought of as analogous to lanes on highways. Specific regions (i.e., lanes) of the spectrum are reserved for specific types of communication (e.g., radio, Wi-Fi). The entire spectrum ranges from 3Hz-300EHz and the typical range used for *wireless communication* today is 30Khz-28GHz.

Spectrum access activity in wireless systems carry rich information which can indicate the presence and activity of physical devices, and behaviors corresponding to security threats, intrusions, jamming attempts, device malfunctioning, interference, illicit transmissions, and a host of other activities (see Fig.1). Data corresponding to spectrum access activity has been explored in wireless intrusion detection systems (WIDS) in a very limited context. Most of the systems in use today for detecting anomalous network activity, are highly application specific and focus on specialized feature engineering, detector engineering, and signal-specific digital signal processing. Such systems are not generalizable, are highly sensitive to minor variations in system characteristics, and are costly to maintain due to the dependence on rich feature engineering.

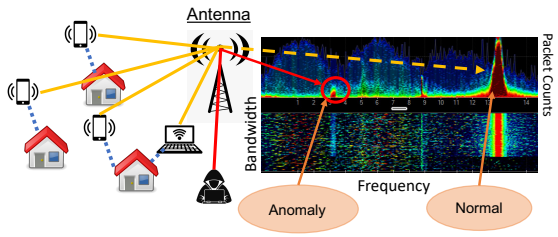

Fig. 1. Irregular Activity in Wireless Communication Systems [1]

Hence in this work, we have developed a generic and powerful unsupervised anomaly detection framework and demonstrated its prowess in the context of wireless network anomalies. Specifically, we propose a novel solution to anomaly detection (AD), *Contrastive Adversarial Anomaly Detection* (CAAD) which applies contrastive learning (CL) in an adversarial setup. We also augment CAAD with the ability to incorporate expert feedback (EF) to improve the quality of its learned representations for AD. We call this model CAAD-EF. To the best of our knowledge, we are the first to propose such a powerful yet flexible AD framework that applies CL paradigms in an adversarial setup with the ability to incorporate expert feedback via CL to improve its learned representations and reduce prediction uncertainty.

Our contributions are as follows:

- We propose CAAD, a novel method for AD which utilizes CL and generative adversarial networks (GAN). We demonstrate that our proposed model is able to significantly outperform state-of-the-art (SOTA) models on AD in wireless networks and standard datasets. To the best of our knowledge,

[1]https://github.com/rgopikrishna-vt/CAAD

CAAD is the first model to use a combination of CL and adversarial learning for AD.

• We propose CAAD-EF, which is another novel model supplemental to CAAD, which further enables us to incorporate expert feedback via CL and uncertainty quantification (using Monte Carlo dropouts). To the best of our knowledge, our framework is the first successful undertaking to utilize CL to incorporate expert feedback.

• We highlight the importance of various facets of CAAD-EF via rigorous qualitative, quantitative, and ablation analyses.

II. Related Work

We now detail some related and recently proposed approaches to anomaly detection that employ deep learning methods. Methods such as autoencoders and GANs [2] have demonstrated state-of-the-art results across many AD tasks. AE models like the Robust Autoencoder [3] (a semi-supervised method requiring a limited set of anomaly labels) and unsupervised AD models based on GANs like AnoGAN [4] and more recently fAnoGAN [5] (employed as a baseline) have shown effective AD performance. Another popular line of AD research leverages contrastive learning. Recently, semi-supervised approaches like *Masked Contrastive Learning* [6] and unsupervised CL approaches like *Contrasting Shifted Instances* (CSI) [7] (employed as a baseline) and *Mean Shifted Contrastive Loss* [8] have also been proposed for AD.

Another line of research proposes AD models that incorporate expert feedback. AAD [9] performs anomaly detection in an interactive data exploration loop with the goal of maximizing the number of anomalous instances presented to the expert (which is different from our setup). Such a method (based on a budget and requiring extensive interactive expert feedback) is intractable for wireless AD where the volume of input data is high. Methods like SAAD [10], DevNet [11], DPLAN [12] and RAMODO [13] are semi-supervised methods, all requiring labeled anomalies and unsuitable for our problem.

In our proposed CAAD-EF framework, we leverage the power of self-supervised contrastive learning and adversarial learning to develop a powerful AD model. We then augment the model's ability to adapt to changing distributional dynamics in AD settings by enabling it to incorporate minimal expert feedback. None of the related AD approaches use a combination of the aforementioned techniques to ensure powerful and robust learned representations for AD.

III. Background

CAAD and CAAD-EF employ techniques such as contrastive learning (CL), generative adversarial networks (GAN), and uncertainty quantification (UQ). We shall now briefly introduce these concepts before detailing the full CAAD-EF framework in section IV.

A. Generative Adversarial Networks (GAN)

We employ the well-studied and stable Wasserstein GAN with gradient penalty (WGAN) model as the backbone of our learning framework. Eq. 1 depicts the WGAN loss function where generator G is parameterized by θ and discriminator D is parameterized by $\Omega \in \mathcal{B}$, where $\mathcal{B}$ is the set of 1-Lipschitz functions.

$$L^{gp} = \min_{\theta} \max_{\Omega \in \mathcal{B}} \mathop{\mathbb{E}}_{x \sim P_r} [D_\Omega(x)] - \mathop{\mathbb{E}}_{\tilde{x} \sim P_f} [D_\Omega(\tilde{x})] + \lambda \mathop{\mathbb{E}}_{\check{x} \sim P_i} (\|\nabla D_\Omega(\check{x})\|_2 - 1)^2 \quad (1)$$

Here, P_r and P_f depicts real and fake distributions and each sample $\check{x} \sim P_i$ is generated as a convex combination of points from P_r, P_f (i.e., sampled from the line connecting points from P_r, P_f). λ enforces the strictness of the gradient penalty.

B. Contrastive Learning (CL)

The paradigm of CL has recently demonstrated highly effective results across a diverse set of disciplines, especially in computer vision [14]–[16]. While most CL losses are set in a *self-supervised* context, a supervised version of the contrastive loss [17] has recently been proposed. A model trained with supervised contrastive learning (SupCon) on a labeled dataset learns latent representations grouped by *class labels* while also forcing separation in representations between instances of different classes.

Consider a dataset of instances $\mathcal{D} = \{(x_1, y_1), .., (x_m, y_m)\}$ such that $x_i \in \mathbb{R}^{b \times l}$ and $y_i \in \mathcal{C}$ is the label of x_i and $\mathcal{C}$ is the set of class labels. Then, the SupCon loss is defined by Eq. 2.

$$L^{sup} = \sum_{x_i \in \mathcal{D}} \frac{-1}{|Pos(i)|} \sum_{x_k \in Pos(i)} \log \frac{\exp(z_i \cdot z_k / \tau)}{\sum_{j \in Q(i)} \exp(z_i \cdot z_j / \tau)} \quad (2)$$

Here, $z_i \in \mathbb{R}^{h \times 1}$ is the latent representation of x_i generated by model M. $\text{Pos}(i) = \{x_k \in \mathcal{D} | y_k == y_i \wedge k \neq i\}$ is the set of instances that form the 'positive set' for x_i. $Q(i) = \{D \backslash x_i\}$. $\tau \in \mathbb{R}^+$ is a hyperparameter. We employ Eq. 2 for CL but with labels generated in a self-supervised manner.

C. Uncertainty Quantification (UQ)

Quantifying decision uncertainty is critical to the success of real-world machine learning (ML) frameworks. It is of special relevance in the current setting of AD wherein the confidence of a model in its decision additionally indicates the urgency of a potential alert issued by the model. While traditional ML models yield point predictions, Bayesian ML provides a framework for capturing model uncertainty. One such UQ approach termed Monte-carlo (MC) dropout [18], entails running a monte-carlo sampling (during inference) of a trained model by randomly masking a set of learned weights of the model each time (i.e., *dropout* [19]). This is akin to sampling from the *approximate posterior* which leads to uncovering the model predictive distribution and hence model decision uncertainty.

IV. Problem Formulation

We now detail our novel methods CAAD and CAAD-EF. Fig. 2 details the overall architecture of CAAD-EF.

A. Self Supervised AD with Negative Transformations

The core of the proposed framework is the Contrastive Adversarial Anomaly Detection (CAAD) model. The structure of the CAAD model resembles a WGAN as described

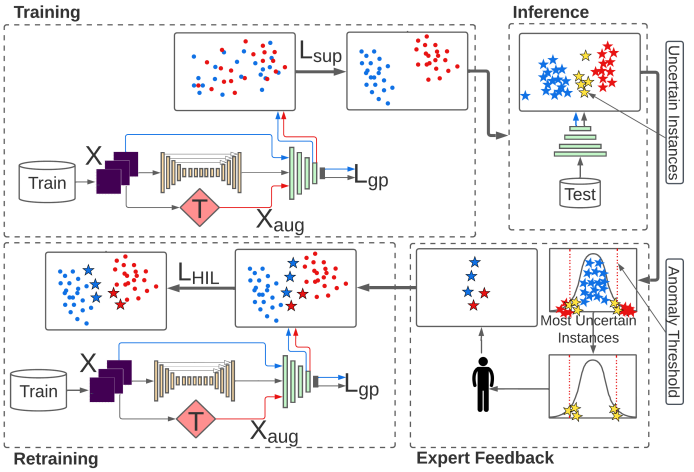

Fig. 2. Full architecture of the human-in-the-loop CAAD-EF AD framework. (Training): The framework consists of a WGAN-GP with an uncertainty-aware discriminator trained with SupCon to impose structure in the latent space. Labeled data required for SupCon is obtained by applying 'negative transformations' on a benign set of instances to generate corresponding anomalous instances. (Inference): During inference, the model yields a prediction (*anomaly*:red star or *benign*:blue star) for every instance, accompanied by the prediction uncertainty. (Expert Feedback): Uncertain instances (yellow stars) are isolated and passed to an *expert* to uncover their true labels. (Re-Training): The pre-trained WGAN-GP model is then fine-tuned with this additional expert feedback to further improve its representations learned thereby leading to improved AD performance and decreased prediction uncertainty.

in section III comprising a generator G_θ and a discriminator D_Ω. In addition to GAN-based training, we also train CAAD discriminator with CL to impose explicit structure on the learned latent representations and improve representation learning. The CL technique employed is similar to SupCon detailed in section III. However, SupCon requires a labeled dataset. To generate a labeled dataset $\mathcal{D}$, we assume the existence of a training set without any anomalies. Let this set be denoted $\mathcal{D}_b = \{(x_1, y_1), .., (x_m, y_m)\}$, such that $y_i = 0, \forall (x_i, y_i) \in \mathcal{D}_b$. We apply a negative transformation $T(\cdot)$ to violate the normalcy of every instance $x_i \in \mathcal{D}_b$ to obtain a corresponding set of anomalous instances $\mathcal{D}_a = \{(x_1, y_1), .., (x_m, y_m)\}$ such that $y_i = 1, \forall (x_i, y_i) \in \mathcal{D}_a$. Now let us consider $\mathcal{D} = \{\mathcal{D}_a, \mathcal{D}_b\} = \{(x_1^a, y_1^a), .., (x_m^a, y_m^a), (x_1^b, y_1^b), .., (x_m^b, y_m^b) | y_i^a = 0 \wedge y_i^b = 1 \, \forall i\}$. Then we can leverage Eq. 2 to directly train the CAAD discriminator with CL (specifically SupCon).

$$L^{CAAD} = L^{gp} + \alpha L^{sup} \tag{3}$$

Eq. 3 represents the objective employed to train the CAAD model to learn effective representations of benign and negatively transformed 'anomalous' instances via CL. Here, α governs the effect of the supervised contrastive loss on the discriminator representations.

B. Inferring Decision Uncertainty with CAAD-UQ

In order to maximize the effect of expert feedback on model performance, we isolate an *effective* set of instances for which we solicit feedback. We define this *effective* set of instances as those for which the model is the most uncertain in its prediction. We augment CAAD to quantify its prediction uncertainty using the popular MC dropout technique (see section III-C). This variant of CAAD augmented with uncertainty quantification capability is termed CAAD-UQ. Concretely, the model structure of CAAD discriminator is augmented by including *dropout* in each layer of the model to yield CAAD-UQ. Let $D^L_{\Omega^*}$ represent the first L layers of the discriminator of a trained CAAD-UQ model where training happens according to Eq. 3. Further, let $d_i = D^L_{\Omega^*}(x_i)$, then, $\{d_i^j\}_{j=1...k}$ represents the set of 'k' monte-carlo sampled embeddings obtained from $D^L_{\Omega^*}(x_i)$. CAAD-UQ employs the mean of the MC embeddings, denoted $\bar{d}_i$ as the representation inferred for an instance $x_i \in \mathcal{D}$.

Every MC embedding d_i^j generated by $D^L_{\Omega^*}(x_i)$ is subjected to a *scoring mechanism* (section IV-D) whereby a prediction $\hat{y}_i^j \in \{0, 1\}$ is obtained. Here $\hat{y}_i^j = 0$ indicates a benign classification and $\hat{y}_i^j = 1$ indicates an anomalous classification of x_i at MC sample j. Prediction uncertainty as quantified by CAAD-UQ for $\bar{d}_i$ is outlined in Eq. 5

$$u_{i,c} = |\{\hat{y}_i^j | j \in \{1, 2, ..k\} \wedge \hat{y}_i^j = c\}| \text{ where } c \in \{0, 1\} \tag{4}$$

$$\mu_i = 1 - \frac{max(u_{i,0}, u_{i,1})}{k} \tag{5}$$

C. Leveraging Expert Feedback

Now, we have formulated a method to calculate an uncertainty measure μ_i for any instance x_i. For $\{x_i\} \in \mathcal{D}$ for which we want to make predictions, if $\mu_i \approx 1 \, \forall x_i$, then we can be sure that the model predictions are reliable. However, if this is not true, we further retrain CAAD-UQ for a small number of epochs using a set of *effective* instances determined using μ_i and their corresponding feedback from an expert on these instances, along with the original training set. We call this model CAAD-EF since this is a contrastive adversarial anomaly detection model trained from expert feedback.

For a particular class c, we define the supervised contrastive loss in Eq. 6.

$$L^{supclass}(\mathcal{D}, c) = L^{sup}(\mathcal{D}) \;\; \forall x_i : y_i = c \tag{6}$$

$L^{supclass}$ is used to only bring instances of one class c together and away from all other classes, in contrast with L^{Sup} which brings each instance close to each other instance of the same class and away from instances of all other classes.

Let X denote a set of benign instances. From the set of inferences yielded by CAAD-UQ, we select the top 'h'% most uncertain instances X^{HIL}, based on μ_i (Eq. 5) as the *effective* set of instances and showcase them to an expert to receive feedback. This feedback gives us $\{X^{HIL}_{anom}, X^{HIL}_{ben}\}$ were X^{HIL}_{anom} and X^{HIL}_{ben} are the set of instances labeled by an expert as anomalous and benign respectively. We then incorporate an additional loss term in the loss function of CAAD-UQ and retrain the model for a small number of epochs. Let $X_{aug} = T(X)$ where T is a class of transformations, $\mathcal{D}_1 = \{X^{HIL}_{anom}, X\}$, $\mathcal{D}_2 = \{X^{HIL}_{ben}, X_{aug}\}$, $\mathcal{D}_3 = \{X^{HIL}_{ben}, X^{HIL}_{anom}\}$. We define the human-in-the-loop (HIL) loss L^{HIL} in Eq 7.

$$L^{HIL} = \alpha_1 L^{supclass}(\mathcal{D}_1, c=1) + \alpha_2 L^{supclass}(\mathcal{D}_2, c=0) + \alpha_3 L^{supclass}(\mathcal{D}_3, c=0) \tag{7}$$

where the first term in L^{HIL} helps bring X^{HIL}_{anom} together while also pushing it far away from X, the second term helps bring X^{HIL}_{ben} together while pushing it far away from X_{aug} and the third term helps bring X^{HIL}_{ben} together and pushes it away from X^{HIL}_{anom}. Hence the overall loss term for the retraining model CAAD-EF is given below.

$$L_D = L^{CAAD} + L^{HIL} \tag{8}$$

D. Anomaly Detection

Training our model gives us meaningful embeddings. Here we define how we use these embeddings to identify anomalies.

Scoring function: We define a scoring function that can be used to determine if an instance is an anomaly or not. We adopt the scoring mechanism used in [7]. Consider a set of training instances. We cluster them into m different clusters and obtain their cluster centroids as $\{x_m\}$. For every test instance x_i, the score is calculated as below.

$$s_{x_i} = max(cosine(D^L_{\Omega^*}(x_i), D^L_{\Omega^*}(x_m))) \; \forall \; x_m \tag{9}$$

Anomaly threshold: Consider a validation set x_v and a distribution P of anomaly scores s_{x_v}.

$$\theta = \arg_\theta \{P(s_{x_v} < \theta) = \phi\} \tag{10}$$

where ϕ is the strictness parameter which can be tuned to control the rate of false positives and false negatives. When s_{x_i} exceeds θ, then we call x_i an anomaly.

V. Experimental Setup

Dataset Description: We consider three wireless emission activity datasets LTW1 , LTW2 , and STW1 as well as the well-known MNIST dataset for evaluation. The wireless emission activity datasets consist of metadata (Bandwidth and Center Frequency values) describing detected radio emissions observed over the air in a known radio frequency (RF) environment. This metadata is aggregated periodically into 80x80 bins based on counts, which are then used as input data. Anomalies consist primarily of new emitters coming online or exhibiting new behavior (e.g. hopping) in a band with otherwise orderly patterned behavior, or the disappearance (e.g. failure) of emitters that are otherwise regularly present. For more information, please refer [20].

Baselines: We evaluated SOTA AD models such as Isolation Forest [21] and OC-SVM [22] as baselines. We also included models which are closely related to CAAD as baselines namely UnetGAN [23], fAnoGAN [5] and CSI [7].

Evaluation Metrics: We use F1score, AUROC, AUPRC, and average weighted F1score as metrics to evaluate our results. The reasoning behind using the metrics can be found in [20].

Model & Training Details: Our models are trained for 100 epochs with a batch size of 32, Adam optimizer, and a learning rate of $1e^{-4}$ for both the generator and the discriminator. We use the penultimate layer in discriminator for L in D^L_Ω. For more details, please refer [20].

VI. Results and Discussion

We now investigate the performance of our models. Our detailed analysis entails a rigorous quantitative and qualitative evaluation. Our specific research questions are as follows:

- How does CAAD perform relative to existing SOTA for AD?
- Can we augment CAAD to incorporate expert feedback (CAAD-EF) to improve the quality of learned representations?
- How does each facet of our novel CAAD-EF framework contribute towards the overall performance?

A. CAAD Anomaly Detection Performance

First, we investigate the AD capability of CAAD . Specifically, we evaluate AD performance across four datasets comprising diverse characteristics and anomalies (see [20]).

TABLE I
Summary of Results.

Data	Model	Benign F1	Anomaly F1	AUROC	AUPRC	Avg.Wt. F1
LTW1	Isolation Forest	0.75	0.47	0.88	0.83	0.61
	OC-SVM	0.41	0.7	0.86	0.81	0.55
	CSI	0.75	0.2	0.61	0.53	0.48
	fAnoGAN	0.69	0.18	0.8	0.8	0.44
	fAnoGAN**	0.68	0.04	0.85	0.84	0.37
	UnetGAN	0.74	0.41	0.86	0.89	0.58
	CAAD	**0.93**	**0.9**	**0.97**	**0.97**	**0.92**
STW1	Isolation Forest	0.64	0	0.49	0.57	0.3
	OC-SVM	0.59	0.79	0.97	0.98	0.7
	CSI	0.03	0.81	0.37	0.58	0.44
	fAnoGAN	0.85	0.72	0.95	0.93	0.78
	fAnoGAN**	0.83	0.79	0.96	0.63	0.81
	UnetGAN	0.85	0.9	**1.0**	**1.0**	0.88
	CAAD	**0.92**	**0.94**	**1.0**	**1.0**	**0.93**
LTW2	Isolation Forest	0.75	0.02	0.63	0.71	0.46
	OC-SVM	0.34	0.59	0.74	0.78	0.44
	CSI	**0.84**	0.27	0.63	0.3	0.61
	fAnoGAN	0.75	0.14	0.7	0.58	0.51
	fAnoGAN**	0.76	0.6	0.73	0.63	0.7
	UnetGAN	0.73	0.36	0.64	0.5	0.58
	CAAD	0.77	**0.73**	**0.86**	**0.83**	**0.75**
MNIST	Isolation Forest	0.28	0.63	0.88	0.59	0.6
	OC-SVM	0.56	0.96	0.91	0.62	0.92
	CSI	0.55	0.9	0.9	0.81	0.87
	fAnoGAN	0.51	0.88	**0.98**	**1.0**	0.84
	fAnoGAN**	0.31	0.65	0.95	0.99	0.62
	UnetGAN	0.5	0.89	0.93	0.99	0.85
	CAAD	**0.76**	**0.97**	0.93	**1.0**	**0.95**

Table I details the AD performance comparison of CAAD with several well-accepted SOTA AD models. Across all the datasets and types of anomalies, CAAD achieves a mean performance improvement of **92.84**% as evidenced by the anomaly *F1 score* metric. CAAD also achieves an overall mean performance improvement of **59.39%** across three of the four datasets where CAAD is the best performing model (i.e., combined performance on benign and AD) as demonstrated by the weighted average F1 score metric.

False Positives: An important facet of a robust, practically useful AD framework is its ability to minimize 'false alarms'. To investigate this behavior, we report AUROC, an explicit function of the false positive rate (FPR). CAAD yields consistently high AUROC values (indicative of its low FPR i.e., it produces very few false alarms). CAAD yields the highest AUROC values in three out of the four datasets. We must note that in the case of the MNIST dataset, the AUROC of CAAD (i.e., **0.93**) is competitive and amenable for use in

real-world AD applications.

Due to the variegated nature of data imbalance in our experiments (see [20] for data support statistics) we also evaluate the AUPRC metric (as a complement to AUROC under data imbalance). We notice that CAAD is best performing[2] across all datasets (including MNIST) as per AUPRC metric.

Network Anomaly Detection: CAAD is able to detect extremely 'weak' anomalies associated with activity in irregular parts of the spectrum being monitored. This is specifically evidenced by the superior performance of CAAD on datasets LTW1 and LTW2, both of which contain attack signatures generated by devices that inappropriately access unused regions of the band being monitored. The superior performance of CAAD in AD on the STW1 dataset which consists of the 'signal drop' anomaly (please refer [20]), also demonstrates the versatility of CAAD to detect different types of irregularities in different bands across the communication spectrum. We notice that the CSI model has a higher benign F1 score but lower overall wt.Avg. F1 score (as it underperforms on the corresponding AD task) for the LTW2 dataset. CAAD in contrast yields more stable results for detecting both benign and anomalous instances across all datasets.

MNIST Anomaly Detection: We notice that CAAD yields a performance improvement of **1.04**% and **3.26**% on F1 score and weighted average F1 score respectively over next best model OC-SVM in the MNIST dataset. This is indicative of generic, flexible nature of CAAD in addressing AD tasks.

SOTA Models: Standard AD models like Isolation Forest, OC-SVM and fAnoGAN are unable to identify the subtle anomalous patterns of interest. CSI which is a recent SOTA AD model that also employs CL, significantly underperforms relative to CAAD (avg. performance improvement by Wt. Avg. F1 score **58.78**%) across all the datasets.

Overall, Table I indicates the superior representation learning and AD capability of CAAD on small and large, balanced and imbalanced datasets with multiple types of anomalies.

TABLE II
IMPACT OF FINE TUNING WITH EXPERT FEEDBACK.

Data	Model	Benign F1	Anomaly F1	AUROC	AUPRC	Avg.Wt. F1
LTW1	CAAD	0.93	0.9	0.97	0.97	0.92
	CAAD-UQ	0.92	0.9	0.97	0.98	0.91
	CAAD-EF	0.94	0.94	0.98	0.98	0.94
	CAAD-EF 95%	0.95	0.94	0.98	0.98	0.95
STW1	CAAD	0.92	0.94	1	1	0.93
	CAAD-UQ	0.93	0.94	1	1	0.94
	CAAD-EF	0.98	0.98	1	1	0.98
	CAAD-EF 95%	0.98	0.99	1	1	0.99

B. Anomaly Detection with Expert Feedback

In an effort to further improve the performance of CAAD, we augment it with the capacity to incorporate expert feedback received in the form of instance labels for a limited set of instances. The resulting framework CAAD-EF comprises of an augmentation to the discriminator of the CAAD model enabling it to characterize its prediction uncertainties (see section IV). This uncertainty-aware model (CAAD-UQ) is trained in a similar fashion to CAAD. Once trained, CAAD-UQ yields inferences on unseen instances accompanied by its prediction uncertainty. We select 'h%' of the most uncertain instances as infered by CAAD-UQ to be labeled by an expert. This labeled set of instances is leveraged in a feedback loop to fine-tune representations learned by CAAD-UQ, thus yielding a holistic HIL AD CAAD-EF framework.

Effect of Expert Feedback (Quantitative Evaluation): Table II shows the results of incorporating expert feedback. In this table, CAAD-EF 95% is CAAD-EF evaluated only on a test set with expert feedback instances removed. Specifically, we notice that incorporating expert feedback on a subset of uncertain instances (we select instances corresponding to 5% of the most uncertain test predictions as candidates for expert feedback) yields an average performance improvement of **4.17**% in Anomaly F1 score and **3.79**% performance improvement in Benign F1 score over the next best model in the case of large (LTW1) and small (STW1) training datasets. This indicates that the CAAD-EF benefits significantly from expert feedback in the context of different anomalies and data sizes. Figures 3a, and b show uncertainty values for 5% of most uncertain instances before retraining (from CAAD-UQ) and after retraining (from CAAD-EF). We clearly notice an improvement in uncertainty scores after retraining. This result further shows improved performance of CAAD-EF.

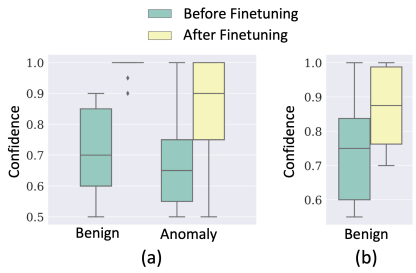

Fig. 3. (a),(b) display uncertainty scores of HIL instances before and after retraining. (a) shows results from LTW1 and (b) shows results from STW1. (b) does not contain values for HIL anomalies as there were no HIL anomalies.

Effect of Expert Feedback (Qualitative Evaluation): To further corroborate our claim of improved representation learning of CAAD-EF due to expert feedback, we analyze the evolution of the discriminator embeddings of CAAD-EF before and after retraining with expert feedback. Fig 4 showcases t-SNE plots of the discriminator embeddings. Fig. 4a shows the representations inferred by CAAD-UQ before expert feedback. We notice clearly the effect of the CL employed to train the discriminator, leading to a clear separation of anomalous and benign regions in the plot. In Fig. 4b we notice CAAD-UQ is uncertain about a significant number of points in the inference set. This region of uncertainty is identified and 5% of most uncertain instances (as indicated by CAAD-UQ) are supplied to the expert for feedback. These instances are highlighted as red (expert label: anomaly) or blue (expert label: benign) points in Fig 4c. The model is retrained with the full training set and the updated sets of points to produce new uncertainty estimates (Fig. 4e) wherein we see that the model is significantly less uncertain in the ROU (which has shrunk significantly). Finally, we notice

[2]accompanied by fAnoGAN on MNIST, UNetGAN on STW1

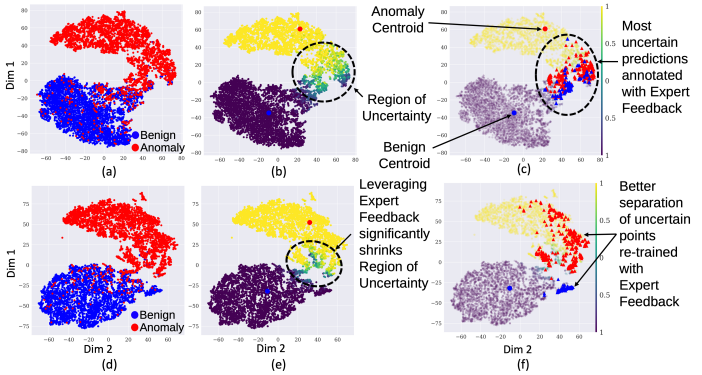

Fig. 4. Fig. 4(a)-(f) qualitatively represent the effect of incorporating human feedback in our proposed CAAD-EF framework. Each figure depicts t-SNE embeddings. Fig. 4(a) Depicts CAAD-UQ embeddings, colored by the ground truth labels of anomalies (red) and benign points (blue). In Fig. 4 (b) the same embeddings are colored by uncertainties obtained from CAAD-UQ (yellow: low uncertainty +, purple: low uncertainty + benign, green regions show uncertain instances;). We notice a highly focused but sizeable 'Region of Uncertainty' (ROU) indicated by the dotted black circle. Fig. 4(c) Shows ground truth labels of ROU points as specified by the expert (red: anomalous, blue: benign). Fig. 4(d) Depicts (similar to Fig. 4a) updated t-SNE embeddings yielded by the CAAD-UQ discriminator after fine-tuning with expert feedback for 5% of most uncertain instances. Fig. 4(e) shows updated uncertainty estimates of CAAD-UQ post fine-tuning, we see a significant reduction in ROU compared to Fig. 4b. In Fig. 4(f) we see embeddings of CAAD-UQ model fine-tuned with expert feedback results in greater separation between benign (blue) and anomalous instances (red) in ROU. This consequently also leads to the overall decrease in decision uncertainty as observed in Fig. 4e.

that the instances that were supplied by the expert as feedback have achieved significant separation and gravitated towards their respective cluster centroids (Fig. 4f) thereby leading to improved model performance. We also performed ablation analysis (please refer [20] for results) to assess the importance of each component of our models. The analysis indicated that the performance of CAAD is a function of the effect of CL and adversarial training, and the performance is further improved with the inclusion of expert feedback (CAAD-EF).

VII. Conclusion

In this paper we have introduced CAAD , a novel AD framework employing contrastive learning in an adversarial setup. We have demonstrated through rigorous experiments that CAAD outperforms SOTA AD baselines and achieves a **92.84**% improvement for AD in wireless communication networks as well as in more generic AD contexts. We further propose CAAD-EF which is a variant of CAAD capable of incorporating expert feedback and evaluated its effectiveness via several qualitative and quantitative experiments. Incorporating expert feedback gives a performance boost of 4.19% over CAAD . Finally, we also highlight the importance of each facet of our proposed CAAD-EF framework through a detailed ablation study. Moving forward, we shall augment CAAD-EF with more sophisticated uncertainty quantification techniques applied to real-time human-in-the-loop AD applications, especially those plagued by covariate shifts.

References

[1] "Wi-fi and non wi-fi interference — metageek," https://www.metageek.com/training/resources/wifi-and-non-wifi-interference/, (Accessed on 10/15/2022).

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[3] C. Zhou and R. C. Paffenroth, "Anomaly detection with robust deep autoencoders," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 665–674.

[4] T. Schlegl *et al.*, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *International conference on information processing in medical imaging*. Springer, 2017, pp. 146–157.

[5] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "f-anogan: Fast unsupervised anomaly detection with generative adversarial networks," *Medical image analysis*, vol. 54, pp. 30–44, 2019.

[6] H. Cho, J. Seol, and S.-g. Lee, "Masked contrastive learning for anomaly detection," *arXiv preprint arXiv:2105.08793*, 2021.

[7] J. Tack, S. Mo, J. Jeong, and J. Shin, "Csi: Novelty detection via contrastive learning on distributionally shifted instances," *Advances in neural information processing systems*, vol. 33, pp. 11 839–11 852, 2020.

[8] T. Reiss and Y. Hoshen, "Mean-shifted contrastive loss for anomaly detection," *arXiv preprint arXiv:2106.03844*, 2021.

[9] S. Das, W.-K. Wong, T. Dietterich, A. Fern, and A. Emmott, "Incorporating expert feedback into active anomaly discovery," in *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, 2016, pp. 853–858.

[10] N. Görnitz, M. Kloft, K. Rieck, and U. Brefeld, "Toward supervised anomaly detection," *Journal of Artificial Intelligence Research*, vol. 46, pp. 235–262, 2013.

[11] G. Pang, C. Shen, and A. van den Hengel, "Deep anomaly detection with deviation networks," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 353–362.

[12] G. Pang, A. van den Hengel, C. Shen, and L. Cao, "Toward deep supervised anomaly detection: Reinforcement learning from partially labeled anomaly data," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 1298–1308.

[13] G. Pang, L. Cao, L. Chen, and H. Liu, "Learning representations of ultrahigh-dimensional data for random distance-based outlier detection," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2018, pp. 2041–2050.

[14] T. Chen *et al.*, "A simple framework for contrastive learning of visual representations," in *ICML*. PMLR, 2020, pp. 1597–1607.

[15] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *CVPR'05*, vol. 1. IEEE, 2005, pp. 539–546.

[16] J. Zbontar *et al.*, "Barlow twins: Self-supervised learning via redundancy reduction," in *ICML*. PMLR, 2021, pp. 12 310–12 320.

[17] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 661–18 673, 2020.

[18] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[19] N. Srivastava *et al.*, "Dropout: a simple way to prevent neural networks from overfitting," *JMLR*, vol. 15, no. 1, pp. 1929–1958, 2014.

[20] G. Rathinavel, N. Muralidhar, T. O'Shea, and N. Ramakrishnan, "Detecting irregular network activity with adversarial learning and expert feedback," *arXiv preprint arXiv:2210.02841v1*, 2022.

[21] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 ICDM*. IEEE, 2008, pp. 413–422.

[22] Y. Wang, J. Wong, and A. Miner, "Anomaly intrusion detection using one class svm," in *Proceedings from the Fifth Annual IEEE SMC Information Assurance Workshop, 2004.* IEEE, 2004, pp. 358–364.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.