Detecting Irregular Network Activity with Adversarial Learning and Expert Feedback

Gopikrishna Rathinavel

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science and Applications

Naren Ramakrishnan, Chair
Tim O'Shea
Chang-Tien Lu
Chandan Reddy

May 6, 2022
Blacksburg, Virginia

Keywords: Anomaly Detection, Network Data Mining, Generative Adversarial Networks

Detecting Irregular Network Activity with Adversarial Learning and Expert Feedback

Gopikrishna Rathinavel

(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 smooth functioning of society. This thesis proposes a novel self-supervised deep learning framework CAAD for anomaly detection in wireless communication systems. Specifically, CAAD employs powerful adversarial learning and contrastive learning techniques to learn effective representations of normal and anomalous behavior in wireless networks. Rigorous performance comparisons of CAAD with several state-of-the-art anomaly detection techniques has been conducted and verified that CAAD yields a mean performance improvement of **92.84%**. Additionally, CAAD is augmented with the ability to systematically incorporate expert feedback through a novel contrastive learning feedback loop to improve the learned representations and thereby reduce prediction uncertainty (CAAD-EF). CAAD-EF is a novel, holistic and widely applicable solution to anomaly detection.

Detecting Irregular Network Activity with Adversarial Learning and Expert Feedback

Gopikrishna Rathinavel

(GENERAL AUDIENCE ABSTRACT)

Anomaly detection is a technique that can be used to detect if there is any abnormal behavior in data. It is a ubiquitous and a 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 smooth functioning of society. Anomaly detection in such communication networks is essential in ensuring security. This thesis proposes a novel framework CAAD for anomaly detection in wireless communication systems. Rigorous performance comparisons of CAAD with several state-of-the-art anomaly detection techniques has been conducted and verified that CAAD yields a mean performance improvement of **92.84%** over state-of-the-art anomaly detection models. Additionally, CAAD is augmented with the ability to incorporate feedback from experts about whether a sample is normal or anomalous through a novel feedback loop (CAAD-EF). CAAD-EF is a novel, holistic and a widely applicable solution to anomaly detection.

Dedication

I would like to dedicate this work to my mom Ms. Saraswathi and my dad Mr. Rathinavel who have constantly supported me. They always inspire me and encourage me to do well. I would like to thank my sister Ms. Kruthika and my brother-in-law Mr. Shankar who have supported me a lot during my masters journey.

Acknowledgments

I would like to thank my advisor and committee chair, Dr. Naren Ramakrishnan for his guidance and support throughout this project. Your passion and spirit is really infectious and encouraged me to work harder and stay focused on my goals.

I am very grateful to Prof. Tim O'Shea for providing me continuous guidance and feedback throughout my research. You have helped me a lot and I am very honoured for having worked with you. I also thank Mr. Mike Piscopo for taking the time to have multiple discussions with me regarding the wireless datasets.

I thank Mr. Nikhil Muralidhar for his invaluable guidance throughout my graduate studies. You were my mentor and guide. I have learnt so much from you and I take those forward with me.

I would like to thank my committee members Dr. Tim O'Shea, Dr. Chandan Reddy and Dr. Chang-Tien Lu for agreeing to serve on my committee. You were very generous with your expertise and precious time.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Wireless systems are pervasive in the modern world, and the radio access network has been expanding in both density and numbers of bands used throughout 4G, 5G and candidate 5G Advanced and 6G systems of the future. These wireless communication systems and networks enable us to access the internet, connect with others remotely, thereby serving as a vital means for human interaction. Further, they connect hundreds or thousands of sensors, applications, industrial networks, critical communications systems, and other infrastructure. Inter-operation of wireless devices across these bands is critical to quality of service and safe operation of communications, broadcast, telemetry, localization, emergency respond and numerous other wireless systems. Therefore, intelligent monitoring of this spectrum access is viewed as one key enabler of the next generation air interface to intelligently coordinate and optimize spectrum among users and applications.

The *electromagnetic spectrum* (simply referred to as 'the spectrum') is the information highway through which most modern 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 the highway. Specific regions (i.e., lanes) of the spectrum are reserved for specific types of communication (e.g., radio communication, broadcast television) based on frequency. The entire spectrum ranges from 3 Hertz (Hz) to 300 EHz. The part used for *wireless communication* ranges from 20 KHz to 300 GHz and the typical range in use today is 30Khz to 28GHz.

Spectrum access activity in wireless systems carries rich information which can indicate underlying activity of physical device presence, activity and behaviors corresponding to security threats and intrusions, jamming attempts, device malfunctioning, interference, illicit transmissions and a host of other activities (see Fig. 1.1). Disruptions or changes to systems within the spectrum today can reflect a wide range of physical world activities, and therefore hold significant value for analysis, safety, optimization, and a range of other applications. Data corresponding to spectrum access activity information has been explored in wireless intrusion detection systems (WIDS) in a very limited context and 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 (DSP) engineering. The widely deployed tools for spectrum sensing focus largely in changes in network level key performance indicators (KPIs) such as BLER, or focus on very narrow applications of RF sensing, such as energy sensing, protocol-specific sensing algorithms, etc. Further, there have been several works [26] [35] looking at ML based sensing of the air-interface, but they largely all rely on high rate radio time-series sample data which is large and cumbersome to store and/or transmit. Such systems are not generalizable, are highly sensitive to minor variations in system characteristics and are costly to maintain due to requirements of rich feature engineering. Another challenge with respect to anomaly detection in wireless signals is that, there is a plethora of possible anomalies and it would be impossible to emulate all such anomalies. Hence, there is a need for an unsupervised model which is capable of detecting any signal that is not normal.

This work focuses on a generic, powerful, scalable and an unsupervised anomaly detection framework which demonstrates its prowess in the context of wireless network anomalies. Specifically, a novel solution to anomaly detection (AD), *Contrastive Adversarial Anomaly Detection* (CAAD) which combines the powerful paradigms of adversarial learning (Gen-

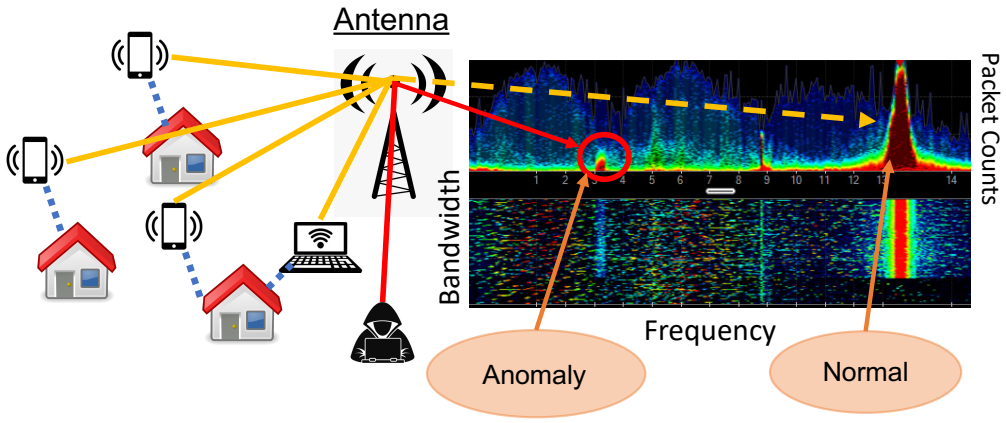

Figure 1.1: Irregular Activity in Wireless Communication Systems.

erative Adversarial Networks - GAN) and contrastive learning (CL) is proposed. CAAD is also augmented with the ability to incorporate expert feedback (EF) to improve the quality of its learned representations for AD. This model is named CAAD-EF (CAAD with expert feedback). To the best of our knowledge we are the first to propose such a powerful yet flexible AD framework that combines contrastive and adversarial learning paradigms with the ability to incorporate expert feedback via contrastive learning to improve its learned representations and reduce prediction uncertainty.

The contributions are as follows:

• CAAD , a novel method for anomaly detection which utilizes generative adversarial networks and contrastive learning is proposed. It is demonstrated that the proposed model is able to significantly outperform state-of-the-art (SOTA) models on anomaly detection in wireless networks and standard datasets. To the best of our knowledge, CAAD is the first model to use a combination of contrastive learning and adversarial learning for anomaly detection.

• Another novel model CAAD-EF , which is supplemental to CAAD is introduced. This model further enables us to incorporate expert feedback via contrastive learning and uncertainty quantification using Monte Carlo dropouts. To the best of our knowledge, CAAD-EF is

the first successful undertaking to utilize contrastive learning to incorporate expert feedback.

• Finally, the importance of various facets of CAAD-EF is highlighted through rigorous qualitative, quantitative and ablation analyses.

Chapter 2

Review of Literature

This chapter reviews literature of some existing work on anomaly detection in the wireless domain and also discusses machine learning methods used for anomaly detection.

2.1 Anomaly Detection in Wireless Signals

The previous works of anomaly detection in wireless signals can be broadly be divided into two segments: 1) anomaly detection on WiFi signals and 2) anomaly detection on raw spectrum data. Since WiFi has become ubiquitous in the current world, there has been lot of interest in exploiting this area. Application specific anomaly detection using WiFi data that have been explored in the recent past includes but not just limited to [42], gesture recognition [1], smoke detection [40], fall detection [36], direction estimation [38]. Literature on using raw spectrum data for anomaly detection is not as extensive as the former. In general some signal processing step is applied to the raw spectrum data before it can be fed into the anomaly detection model. Examples of such data include spectrogram of time-frequency or spectral correlation function outputs [21, 34], power spectral density data [25, 27] and dynamic spectrum access data [15, 20].

Although existing literature addresses anomaly detection in wireless data, they are either storage or compute intensive, are very application specific, or do not provide means to provide expert feedback so that new information can be incorporated into the models. The method

proposed in this thesis uses metadata as opposed to the raw RF signals which immensely reduces storage and compute. Due to this, the proposed method can be used for anomaly detection in various applications. Also, as opposed to the existing literature, the proposed human-in-the-loop framework CAAD-EF , provides a means to incorporate expert feedback into the model. The proposed methods can also be extended outside the scope of wireless data and to support this claim some results on MNIST data are presented.

2.2 Machine Learning for Anomaly Detection

Many machine learning approaches have been developed for anomaly detection across diverse applications. The recent resurgence of deep learning techniques demonstrating their effectiveness across a wide variety of domains has lead to the development of many novel and powerful modeling paradigms like generative adversarial networks (GAN) [11], self-supervised representation learning [17] and contrastive learning (CL) [7].

Contrastive Learning (CL) imposes structure on the latent space by encouraging similarity in representations learned for *related* instances and dissimilarity in representations for unrelated instances. Such techniques have proven effective, especially when combined with self-supervised learning [3, 16] and also with labeled data [18]. CL has demonstrated promising results in image recognition tasks. However, all these efforts have mostly focused on improving representation learning performance on traditional classification tasks and do not specifically focus on anomaly detection. *Generative Adversarial Networks (GANs)* [11] are a powerful generative learning paradigm grounded in an adversarial training setup but fraught with training instability. Recently, improvements have been proposed to stabilize the GAN training setup by employing *Wasserstien* distance functions [2] and gradient penalties on the learned weights.

2.2.1 Deep Learning for Anomaly Detection

The aforementioned developments of the CL, GAN based paradigms have led to such techniques being employed for the ubiquitous and challenging problem of anomaly detection. Specifically, in [41], a deep robust autoencoder (Robust AE) model is proposed, inspired by Robust Principal Component Analysis technique, for anomaly detection with noisy training data. However, this methodology by design requires knowledge of a subset of anomalies during model training and may be considered semi-supervised and is not directly related to our context of unsupervised AD. Recently, another line of anomaly detection research [30] proposes employing DCGAN [11] for unsupervised anomaly detection. The authors then build upon their previous work to propose fAnoGAN [31], a two step encoder-decoder architecture based on DCGANs where the encoder (trained separately) learns to invert the mapping learned by the Generator (i.e., decoder) of the DCGAN model. fAnoGAN is employed as one of the baselines for empirical comparison.

2.2.2 Contrastive Learning for Anomaly Detection

There are multiple reports of contrastive learning being utilzed for AD. *Masked Contrastive Learning* [5] is a supervised method that varies weights of different classes in the contrastive loss function to produce good representations that separate each class. Even though this method shows promise, it requires knowledge of anomaly labels. *Contrasting Shifted Instances* (CSI) [33] and *Mean Shifted Contrastive Loss* [28] are two unsupervised AD methods based on CL. CSI investigates the power of self-supervised CL for detecting out-of-distribution (OOD) data by using distributionally shifted variations of input data. CSI is employed as one of the baselines. *Mean Shifted Contrastive Loss* applies a contrastive loss modified using the mean representation on representations generated using models pre-

trained on ImageNet data. However, this model is not useful for wireless AD as it is pre-trained on a particular kind of data. Also, none of these methods provide a means to incorporate expert feedback.

2.2.3 Incorporating Expert Feedback

The solutions presented in [8, 12, 22, 23, 24] all employ human feedback in various ways. Active Anomaly Discovery (AAD) [8] is designed to operate in an anomaly exploration loop where the algorithm selects data to be presented to experts and also provides a means to incorporate feedback into the model. However, its performance is dependent on the number of feedback loops that can be afforded. Hence, such a method could not be applied for wireless anomaly detection where the volume of input data is really high. RAMODO [24], combines representation learning and outlier detection in a single objective function. It utilizes pseudo labels generated by other state-of-the-art outlier detection methods and Chebyshev's inequality. This dependence on other methods to generate pseudo labels can sometimes be unreliable in cases where state-of-the-art outlier detection methods perform poorly. SAAD [12], DevNet [22] and DPLAN [23] are semi-supervised methods, all of which require minimal labelled anomalies and is not suitable for our problem.

The advantage of using contrastive learning for anomaly detection is that it can be utilized in a self-supervised setup. Hence, the training samples can be augmented to generate anomalous samples that are very close to the training distribution and utilize them as negative samples in the contrastive loss. Also, the penultimate layer of the GAN discriminators have recently been shown to act as good representations of the input data [4, 6, 14, 31]. Hence, the combination of these powerful techniques, CL and GAN serve well for our AD task. None

of the related approaches outlined above have developed AD techniques that combine the aforementioned techniques for anomaly detection. Also, none of the state-of-the-art related AD approaches provide a means to incorporate expert feedback via contrastive learning.

Chapter 3

Background

CAAD-EF employs adversarial learning, contrastive learning (CL) and uncertainty quantification (UQ) techniques. This chapter introduces these concepts before detailing the full CAAD-EF framework in chapter 4.4.

3.1 Generative Adversarial Networks (GAN)

GANs are a class of generative models where the learning problem is formulated as a game between two neural networks, namely the *generator* (G) and the *discriminator* (D). The problem setup of GANs comprises of the generator learning to transform inputs sampled from a noise distribution into a distribution P_f such that it resembles the true data distribution P_r. Essentially, the generator is trained to *fool* the discriminator while the discriminator is tasked with distinguishing between *fake* samples $\tilde{x} \sim P_f$ generated by G and *real* samples $x \sim P_r$. The traditional GAN [11] setup minimizes the Jenson-Shannon (JS) divergence between P_r and P_f. However, this divergence is not continuous with respect to the parameters of G, leading to training instabilities. Wasserstein GAN [2] (WGAN) was proposed to address this issue. WGANs employ the Earth-Mover distance (instead of the JS divergence) which under mild assumptions does not have discontinuities and is almost universally differentiable. Consider the discriminator (also termed the *critic*[1] in WGANs) D parameterized by Ω and

[1] words 'critic', 'discriminator' are used interchangeably in the paper.

generator G parameterized by θ. Eq. 3.1 depicts the WGAN loss function where D is parameterized by $\Omega \in \mathcal{B}$, where $\mathcal{B}$ is the set of 1-Lipschitz functions.

$$L^w = \min_{\theta} \max_{\Omega \in \mathcal{B}} \mathop{\mathbb{E}}_{x \sim P_r} [D_\Omega(x)] - \mathop{\mathbb{E}}_{\tilde{x} \sim P_f} [D_\Omega(\tilde{x})] \tag{3.1}$$

Enforcing the 1-Lipschitz constraint on D_Ω has been found to be challenging. On the basis of the property that a function is 1-Lipschitz if and only if it has norm no greater than 1 everywhere, [13] proposed a solution of augmenting the WGAN loss with a soft-constraint enforcing that the norm of the discriminator gradients (w.r.t the inputs) be 1. The objective function of the WGAN with this updated soft-constraint (termed a gradient penalty) is shown in Eq. 3.2.

$$L^{gp} = L^w + \lambda \mathop{\mathbb{E}}_{\check{x} \sim P_i} \left(\|\nabla D_\Omega(\check{x})\|_2 - 1 \right)^2 \tag{3.2}$$

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 strictness of the gradient penalty.

3.2 Contrastive Learning (CL)

The paradigm of contrastive learning (CL) has recently demonstrated highly effective results across a diverse set of disciplines and tasks, especially in computer vision [3, 7, 39]. The goal of CL is to impose structure on latent representations learnt by a model (M). This is often achieved using soft-penalties (e.g., additional loss terms) that influence representations generated by M to be structured so representations of *related* instances are closer together relative to instances that are known to be *unrelated*. Most CL losses are set in a *self-supervised* context where *relatedness* is generated via augmentations of an instance and two distinct instances are considered to be unrelated.

Recently, [18] proposed *supervised contrastive learning* (SupCon), which is an extension of the CL paradigm to supervised (classification) settings. A model trained with SupCon on a labelled dataset learns latent representations grouped by *class labels* while also forcing separation in representations between instances belonging to different classes (i.e., low intra-class separation and high inter-class separation of latent representations).

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 supervised contrastive loss is defined by Eq. 3.3.

$$L^{sup} = \sum_{x_i \in \mathcal{D}} \frac{-1}{|Pos(i)|} \sum_{x_k \in Pos(i)} \log \frac{\exp\left(z_i \cdot z_k / \tau\right)}{\sum_{j \in Q(i)} \exp\left(z_i \cdot z_j / \tau\right)} \tag{3.3}$$

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. Eq. 3.3 is employed for CL but with labels generated in a self-supervised manner.

3.3 Uncertainity 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 anomaly detection 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 [10], can be considered to approximate Gaussian Processes with neural network models. This approach termed Monte-Carlo Dropout 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* [32]). This is akin to sampling from the *approximate posterior* which leads to uncovering the model predictive distribution. Inferring the predictive distribution is one of the methods to quantify model uncertainty with Bayesian neural networks.

Chapter 4

Problem Formulation

This chapter describes the various facets of our novel human-in-the-loop anomaly detection framework CAAD-EF . Fig. 4.1 details the overall architecture of CAAD-EF .

4.1 Self Supervised Anomaly Detection 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 with gradient penalty (WGAN-GP) as described in chapter 3 comprising a generator G_θ and a discriminator D_Ω. In addition to GAN based training, CAAD discriminator is also trained with CL to impose explicit structure on the learned latent representations and improve representation learning. The CL technique employed is similar to supervised contrastive learning (SupCon) detailed in chapter 3. However, SupCon requires a labeled dataset. To generate a labeled dataset $\mathcal{D}$, existence of a training set without any anomalies is assumed. Let this set be denoted $\mathcal{D}_b = \{(x_1, y_1), .., (x_m, y_m)\}$, such that $y_i = 0$, $\forall (x_i, y_i) \in \mathcal{D}_b$. A negative transformation $T(\cdot)$ is applied 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 $\mathcal{D} = \{\mathcal{D}_a, \mathcal{D}_b\} = \{(x_1^a, y_1^a), .., (x_m^a, y_m^a), (x_1^b, y_1^b), .., (x_m^b, y_m^b) | y_i^a = 0 \wedge y_i^b =$

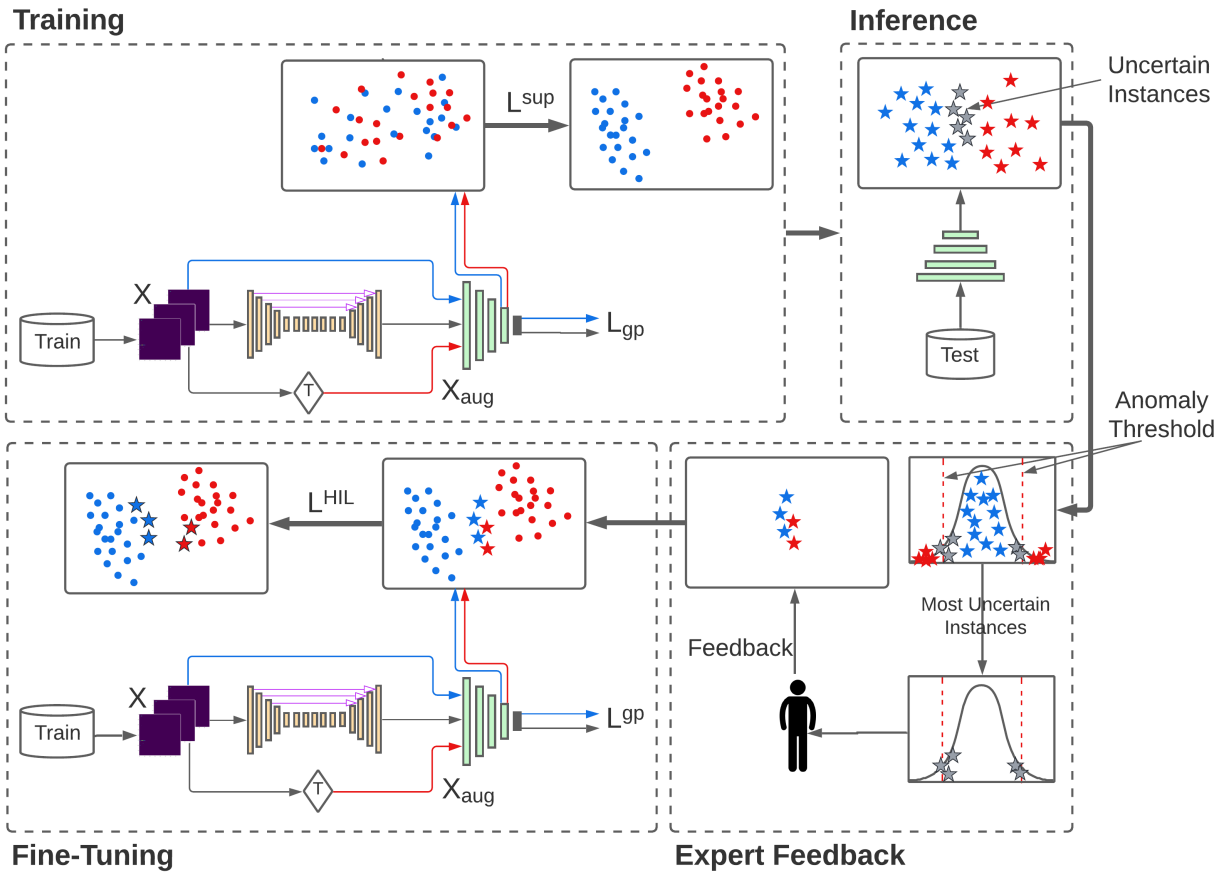

Figure 4.1: **Full architecture of the human-in-the-loop CAAD-EF anomaly detection framework**. (Training): The framework consists of a WGAN-GP with an uncertainty-aware discriminator trained with supervised contrastive learning (SupCon) to impose structure in the latent space. Labeled data required 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 (grey stars) are isolated and passed to an *expert* to uncover their true labels. (Fine-Tuning): 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 anomaly detection performance and decreased prediction uncertainty.

$1\,\forall i\}$. Eq. 3.3 is then leveraged to directly train the CAAD discriminator with CL (specifically SupCon).

$$L^{CAAD} = L^{gp} + \alpha L^{sup} \tag{4.1}$$

Eq. 4.1 represents the objective employed to train the CAAD model to learning 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.

4.2 Inferring Decision Uncertainty with CAAD-UQ

In order to maximize the effect of expert feedback on model performance, *effective* set of instances should be isolated to solicit feedback. To this end, *effective* set of instances are defined as those for which the model is the most uncertain in its prediction. Hence CAAD, trained to learn effective representations as defined in section 4.1, is augmented to quantify its prediction uncertainty using the popular Monte-Carlo dropout technique (see section 3.3). 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. 4.1. 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 Monte-Carlo embeddings, denoted $\overline{d}_i$ as the representation inferred for an instance $x_i \in \mathcal{D}$.

Every Monte-Carlo embedding d_i^j generated by $D^L_{\Omega^*}(x_i)$ is subjected to a *scoring mechanism* (section 4.4) 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 $\overline{d}_i$ is outlined in Eq. 4.3

$$u_{i,c} = |\{\hat{y}_i^j | j \in \{1, 2, ..k\} \wedge \hat{y}_i^j = c\}| \text{ where } c \in \{0, 1\} \tag{4.2}$$

$$\mu_i = 1 - \frac{max(u_{i,0}, u_{i,1})}{k} \tag{4.3}$$

4.3 Leveraging Expert Feedback

Now, a method to calculate an uncertainty measure μ_i for any instance x_i is formulated. For $\{x_i\} \in \mathcal{D}$ for which predictions has to be made, if $\mu_i \approx 1 \ \forall \mathrm{x}_i$, then the model predictions are reliable. However, if this is not true, CAAD-UQ is further finetuned using a set of *effective* instances determined using μ_i and their corresponding feedback from an expert on these instances. This model is called CAAD-EF since this is a contrastive adversarial anomaly detection model trained from expert feedback.

For a particular class c, the supervised contrastive loss is defined in Eq. 4.4.

$$L^{supclass}(\mathcal{D}, c) = L^{sup}(\mathcal{D}) \ \ \forall x_i : y_i = c \tag{4.4}$$

$L^{supclass}$ 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, the top 'h'% most uncertain instances are selected X^{HIL}, based on μ_i (Eq. 4.3) as the *effective* set of instances and showcase them to an expert to receive feedback. This feedback gives $\{X^{HIL}_{anom}, X^{HIL}_{ben}\}$ where X^{HIL}_{anom} and X^{HIL}_{ben} are the set of instances labeled by expert as anomalous and benign respectively. An additional loss term is then incorporated in the loss function of CAAD-UQ and finetune the model. Let $X_{aug} = T(X)$ where T is a class of transformations, $\mathcal{D}_1 = \{X^{HIL}_{anom}, X\}$, $\mathcal{D}_2 = \{X^{HIL}_{ben}, X_{aug}\}$, $\mathcal{D}_3 = \{X^{HIL}_{ben}, X^{HIL}_{anom}\}$. The HIL

loss L^{HIL} is defined in Eq 4.5.

$$\begin{aligned} L^{HIL} &= \alpha_1 L^{supclass}(\mathcal{D}_1, c=1) + \alpha_2 L^{supclass}(\mathcal{D}_2, c=0) \\ &+ \alpha_3 L^{supclass}(\mathcal{D}_3, c=0) \end{aligned} \tag{4.5}$$

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 finetuning model CAAD-EF is given below.

$$L_D = L^{CAAD} + L^{HIL} \tag{4.6}$$

4.4 Anomaly Detection

Now that we have meaningful embeddings after training the model, we need a method to determine if a given instance is benign or anomalous. This section explains how embeddings are used for identifying anomalies.

Scoring function: Since we are employing contrastive loss to create a separation between benign embeddings and embeddings of anomalous augmentations, we would be able to benefit from a scoring function that utilizes the distance measure which is used in the contrastive loss. This distance measure is the *Cosine Distance*. The scoring mechanism we have used that uses cosine distance is adopted from [33]. Consider a set of instances used during training. They are clustered into m different clusters and their cluster centroids are obtained as $\{x_m\}$.

For every test instance x_i, the score is calculated as below.

$$s_{x_i} = max(cosine(D^L_{\Omega^*}(x_i), D^L_{\Omega^*}(x_m))) \; \forall \; x_m \tag{4.7}$$

This method of creating clusters of training embeddings and then determining the maximum cosine distance from the test instance, creates a more refined boundary between the benign embeddings and possible anomalous embeddings.

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

$$\theta = \arg_\theta\{P(s_{x_v} < \theta) = \phi\} \tag{4.8}$$

where, when s_{x_i} exceeds θ, then x_i is called an anomaly. ϕ is the strictness parameter which can be tuned to control the rate of misclassifications. When this parameter is low, there will be false positives and when it is high, there will be false negatives. Hence, an apt value should be chosen for the strictness parameter to balance the number of false positives and false negatives.

Chapter 5

Experimental Setup

Several wireless emission activity datasets as well as the well-known MNIST computer vision dataset are considered for evaluation. The wireless emission activity datasets consist of metadata describing detected radio emissions observed over the air in a known radio frequency (RF) environment. Metadata describes a range of aspects including detection time, frequency, bandwidth, signal type, signal power, and is streamed from an edge sensor into an Elasticsearch database for archival. 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 which are otherwise regularly present.

5.1 Dataset Description

Using Software Defined Radio (SDR), metadata particular to FM signals were gathered. This metadata is in the form of JSON with each instance containing information about packets of signals and a timestamp. Centre frequency and bandwidth are the two useful features extracted from the metadata. 80x80 bins are created with bandwidth on x-axis and frequency on y-axis and populate the count of packets falling into each bin for a period of 3 minutes. This gives us a time series of 80x80 images. For context, the maximum count of packets per pixel in the LTW1 dataset is 98 and maximum count of packets per pixel in

the LTW2 dataset is 201. Also, mean of maximum count of packets per pixel per image is 20.5 for LTW1 and 107.1 for LTW2 . This indicates that LTW2 is more dense compared to LTW1 . Min-max normalization based on the global min and max of the count of packets per pixel of training set is performed. Once such normalized images are gathered, the training set is denoised by masking out pixels that have a probability of having a non-zero value of 0.0005. Below is the detailed description of individual datasets.

LTW1 : Long-term wireless emission metadata in the 800-900MHz band for a span of 2 months between 8 Apr 2021 to 8 Jun 2021 from a first location with a frequency scanning receiver. 11670, 5002, 7146 are the number of training, validation and test sets respectively. The test set contains 3894 benign and 3738 anomalies which also includes hopping anomalies.

LTW2 : Long-term wireless emission metadata (similar to LTW1) from a second geographical location. Data spans a time interval of 24 Nov 2020 to 4 Dec 2020. The test set contains 786 benign instances and 724 anomalies.

STW1 : Short-term wireless emission metadata comprising short-time high-rate observations of 900MHz cellular and ISM bands. Here, input is similar to 80x80 bin density features of LTW1 but with a 1-second interval. 154, 39, 198 are the number of training, validation and testing instances respectively. The test set contains of 94 benign instances and 104 anomalies.

MNIST: This is a standard image dataset [9] for machine learning research comprising images of hand-written digits. The training and validation set consists of 4089 and 1753 images of the number '4' (benign class) respectively. During testing, all other classes of digits are considered anomalies. The test set contains 982 benign and 9018 anomalous images leading to an imbalanced evaluation setup. We view our anomaly detection framework as one which can be extended outside of the wireless anomaly detection setup. Hence, we test

our model on MNIST to verify generalisability.

5.2 Types of Wireless Anomalies

1. **Hopper Anomalies:** In this type of anomaly, frequency hopper signals are transmitted. This creates activity in different regions of the 80x80 image. Examples of such hopper anomalies are shown in Figures 5.1a and 5.1b.

2. **Drop Anomalies:** In this type of anomaly, on LTE signal in the cellular band goes offline at around 198 seconds into the dataset. Data until 170 seconds was used for training and the rest, with anomalies for testing. This removes a region of activity from the 80x80 image. Examples of a drop anomaly is shown in Figure 5.1c.

3. **Naturally occurring anomalies:** In the training set, as mentioned in Section 5.1, pixels that have a probability of having a non-zero value of 0.0001 were masked out. However this is not done in testing set and such images are instead labelled as anomalies.

5.3 Anomaly Injection

1. **Injection of hopper anomalies:** The available number of real world hopper anomalies was 60 packets with different bandwidth and frequencies for each packet. These 60 packets were then grouped into numbers of 6 which effectively produced 10 different anomaly signatures. Now, every anomaly signature contains 6 anomalies with each having a count (density) of 1.

   These anomaly signatures were used as template anomalies and injected (added) them

in test instances without anomalies while also keeping a copy of original test instance without anomalies. Before anomaly injection, the template is min-max normalized using the global min and max of the count of packets per pixel of the entire training set. As LTW2 has max of the count of packets per pixel of the entire training set value as 201 and since our anomalies have a density of 1, during min-max normalization, the injected anomalies become very subtle. This is one explanation for the relatively low anomaly F1 scores in LTW2 compared to other datasets. Figure 5.1b shows anomalies amplified by 30x in-order to be visible to the naked eye.

2. **Injection of Drop Anomalies:** To inject drop anomalies, LTE signal is manually turned off.

5.4 Baselines

Isolation Forest [19]: A popular bench mark ensemble method for anomaly detection wherein partitions are created in data such that each data point is isolated. During such partitioning, an anomalous data instance isolates itself much easier compared to a benign instance. 100 base estimators were employed.

One Class SVM (OC-SVM) [37]: Another popular anomaly detection benchmark, OC-SVM, learns a hypersphere that encompasses points from a single class. Any point that falls outside of this hypersphere is considered an outlier. The kernel we have used in the algorithms is a radial basis function kernel.

UNetGAN [29]: GAN model with UNet as generator, and convolutional network used for the discriminator. The GAN model is trained using Wasserstien loss and gradient penalty.

fAnoGAN[31]: A state-of-the-art anomaly detection model based on Wasserstien GAN-GP. Specifically, fAnoGAN comprises a two stage-anomaly detection framework wherein the first stage involves WGAN-GP being trained to translate samples from a noise distribution to outputs of the process of interest. The second stage employs the pre-trained WGAN-GP to train an encoder network to learn an 'inverse mapping'of WGAN-GP generator. The anomaly score is calculated as a function of error between intermediate representations of input data and generator output. This model can serve as a good baseline because it is similar to our model setup in two ways: 1) this model uses a WGAN for anomaly detection and 2) this model uses the penultimate layer of the discriminator for anomaly scoring.

Contrasting Shifted Instances (CSI)[33]: A state-of-the-art anomaly detection deep neural network models which employs traditional self-supervised contrastive learning with a novel contrastive task comparing an instance to *distributionally shifted* versions of itself. Anomaly detection occurs through a novel anomaly scoring mechanism. This model serves as a good baseline as it employs contrastive learning in the anomaly detection task.

5.5 Evaluation Metrics

Multiple evaluation metrics were used to provide a holistic quantitative evaluation of model performance on the task of anomaly detection. Specifically, *F1 scores* were reported for correctly detecting benign and anomaly instances as well as a *weighted average* of the two F1 scores. Further, Area Under the Receiver Operating Characteristic (AUROC) metric which is explicitly dependent on the false positive rates (FPR) of the models are also reported. AUROC is explicitly reported to investigate whether models have low FPRs (thereby high AUROC) values which is imperative for an effective anomaly detection model. Finally, as imbalanced datasets (see section 5.1) are also used, the Area Under the Precision-Recall

Curve (AUPRC) metric which is known to complement AUROC well by alleviating biases due to data imbalance is utilized.

5.6 Model & Training Details

Model details: CAAD-EF is a WGAN with gradient penalty (WGAN-GP). The generator of this WGAN-GP is a UNet autoencoder comprising 5 down convolutional layers (each with kernel size 4, batch normalization and leaky-ReLU), 5 same convolutional layers (each with kernel size 3, maxpooling, batch normalization and leaky-ReLU) and 5 up-convolutional layers (each with kernel size 4 batch normalization and ReLU). Discriminator has five convolutional layers, each followed by instance normalization, leaky-ReLU (negative slope of 0.2), dropout layer with dropout probability 0.5.

Training details: Our models are trained for 100 epochs with batch size of 32, Adam optimizer ($\beta 1 = 0, \beta 2 = 0.9$), learning rate of $1e^{-4}$ for both the generator and the discriminator. For gradient penalty, λ value of 10 is used. The weighting coefficients $\alpha, \alpha_1, \alpha_2$ and α_3 are set to 1 during training. The negative transformation that is employed is *salt noise* for the network datasets (LTW1 ,LTW2 ,STW1) and rotations for MNIST. Strictness parameter $\phi = 0.99$ and $m = 1$ are used during validation. A h value of 5 (corresponding to 5% of the most uncertain instances) is used during finetuning.

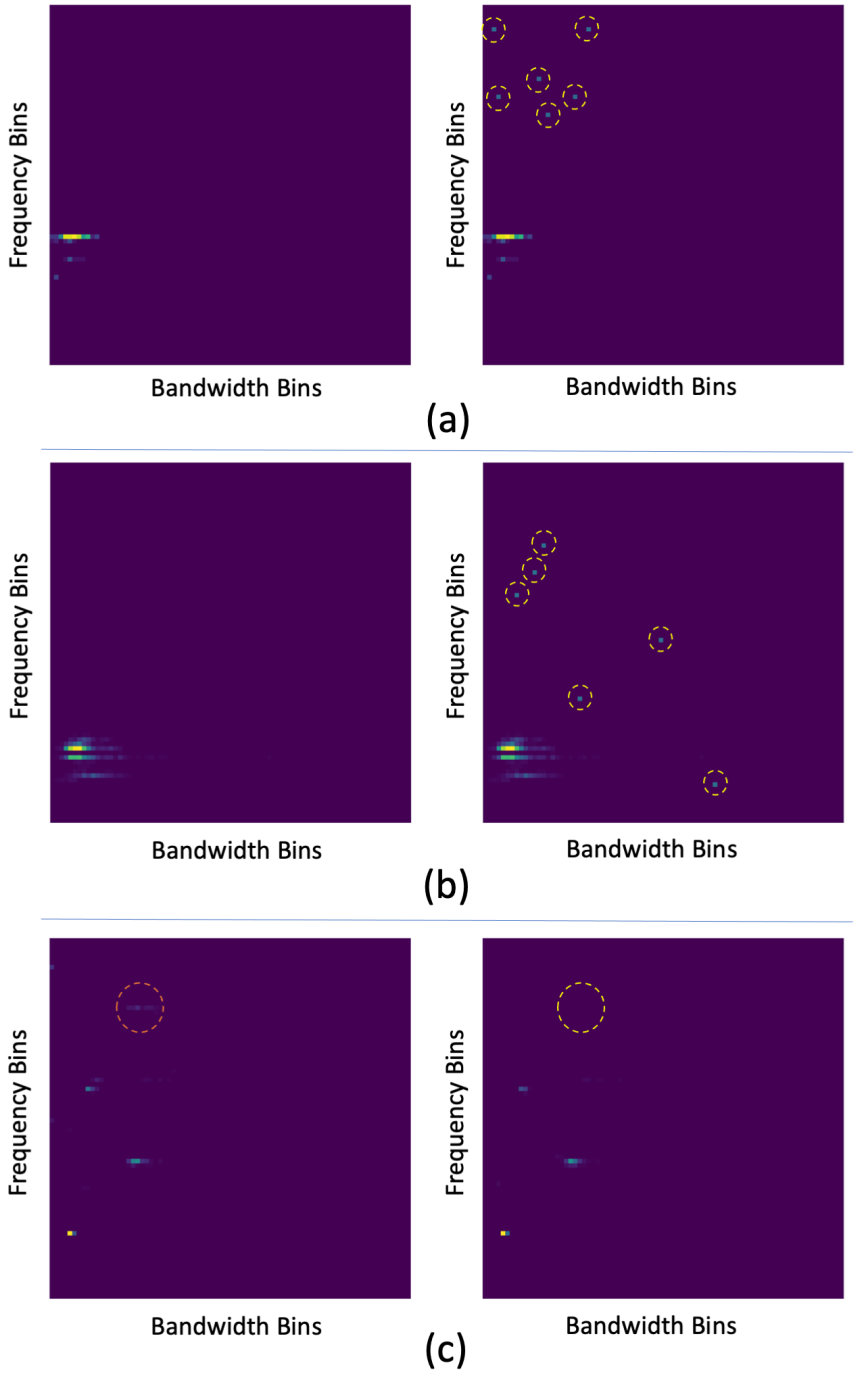

Figure 5.1: **Preprocessed wireless emission activity data**. These are examples of preprocessed wireless emission activity data that are used as inputs to the models. Left figures indicate benign instances and right figures indicate corresponding anomalous instances. Anomalies are annotated with dashed yellow circles. Figure (a) is a sample from the LTW1 dataset. Anomalies in the right figure of (a) are 7x amplified in order to improve visibility. Figure (b) is a sample from the LTW2 dataset. Anomalies in the right figure of (b) are 30x amplified in order to improve visibility. Figure (c) is a sample from the STW1 dataset. In the left figure in (c), the orange circle points to the signal that goes missing in the right figure (Drop anomaly).

Chapter 6

Results and Discussion

The performance of our novel CAAD-EF anomaly detection framework is investigated in this chapter. A detailed analysis that is performed, entails a rigorous quantitative and qualitative performance evaluation. The specific research questions that are asked are as follows:

• How does our CAAD model perform relative to the existing state-of-the-art (SOTA) for anomaly detection?

• Can CAAD be augmented to successfully 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?

6.1 CAAD Anomaly Detection Performance

At the outset, the anomaly detection capability of the CAAD model which forms the backbone of our proposed, novel CAAD-EF framework is evaluated. Specifically, anomaly detection performance of CAAD across four datasets comprising diverse characteristics and anomalies (see section 5.1) is evaluated.

Table 6.1 details the anomaly detection performance comparison of CAAD with several well accepted state-of-the-art (SOTA) anomaly detection models. Across all the datasets and

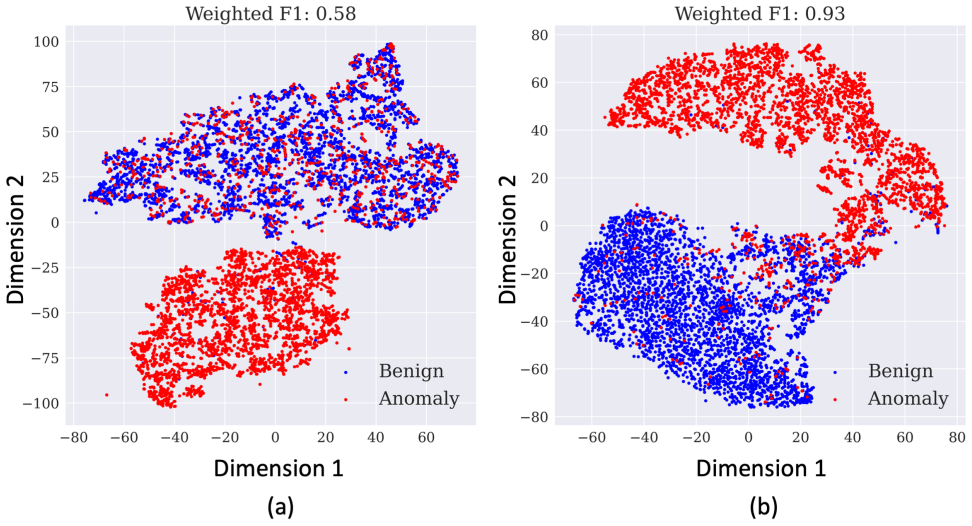

Figure 6.1: **Effect of contrastive loss**. These figures qualitatively represent the effect of contrastive loss in CAAD on the LTW1 dataset. They depict the t-SNE embeddings generated by the discriminator. Figure (a) shows results of CAAD w/o contrastive loss and figure (b) shows results of CAAD with contrastive loss. Figure (b) clearly shows better separation between bening and anomalous embeddings.

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 anomaly detection) as demonstrated by the weighted average F1 score metric.

False Positives: An important facet of a robust and practically useful anomaly detection framework is its ability to minimize 'false alarms'. To investigate this behavior, AUROC values (see 5.5), which explicitly is a function of the false positive rate (FPR) are reported.

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. It must be noted 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 section 5.1 for data support statistics) the AUPRC metric (as a complement to AUROC under data imbalance) is also evaluated. CAAD is the best performing[1] across all datasets (including MNIST) as per the AUPRC metric.

Finally, the effect of contrastive loss in CAAD can be seen in Figure. 6.1 where the embeddings of benign and anomalous samples are more separated when contrastive loss is added.

6.1.1 Network Anomaly Detection

CAAD is able to detect extremely 'weak' anomalous signatures 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 anomaly detection on the STW1 dataset which consists of the 'signal drop' anomaly (described in section 5.1), also demonstrates the versatility of the CAAD model to detect different types of irregularities in different bands across the communication spectrum. The CSI model has a higher benign F1 score but lower overall wt.Avg. F1 score (as it under performs on the corresponding anomaly detection task) for the LTW2 dataset. CAAD in contrast yields more stable results for detecting both benign and anomalous instances across all datasets.

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

6.1.2 MNIST Anomaly Detection

Once again CAAD is able to outperform all other SOTA models for anomaly detection on the MNIST dataset. This result is significant, as it is indicative of the generic nature and flexibility of the proposed solution in addressing variegated anomaly detection tasks. CAAD yields the best anomaly detection F1 score (**1.04**% improvement over next best model OC-SVM) and the best weighted average F1 score (**3.26**% improvement over next best model OC-SVM). CAAD is also the best performing for the benign instance recognition (indicated by benign F1 score). The inconsistency in model performance across the benign and anomaly detection tasks is once again evident in the context of skewed results in the OC-SVM (fails to detect benign instances accurately). Hence, as evidenced by the Wt. Avg. F1 score, Benign F1 score and Anomaly F1 score CAAD yields the best performance on the MNIST anomaly detection task.

6.1.3 SOTA Models

Standard anomaly detection models like Isolation Forest and OC-SVM perform poorly especially for the challenging *network anomaly detection* task. They are unable to identify the subtle anomalous patterns of interest. The fAnoGAN variants showcase unstable performance across the two tasks of benign and anomaly detection as evidenced by the large differences in the F1 scores for each dataset thereby rendering them practically ineffective for use as real-world anomaly detection frameworks. CSI which is a recent SOTA anomaly detection model that also employs contrastive learning, significantly under performs relative to CAAD (avg. performance improvement by Wt. Avg. F1 score **58.78**%) across all the datasets.

Overall, experiments in Table 6.1 indicate the superior representation learning and anomaly

detection capability of CAAD on small and large, balanced and imbalanced datasets comprising multiple different types of anomalies.

6.2 Anomaly Detection with Expert Feedback

Real-world systems can often benefit from incorporating valuable expert knowledge to influence their representation learning capabilities. In an effort to further improve the performance of CAAD , it is augmented 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 4.4). 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. 'h%' of the most uncertain instances are selected as inferred by CAAD-UQ to be labeled by an expert. This labeled set of instances is leveraged in a feedback loop to fine-tune the representations learned by CAAD-UQ , thus yielding the holistic human-in-the-loop anomaly detection CAAD-EF framework.

6.2.1 Effect of Expert Feedback (Quantitative Evaluation)

To investigate the effectiveness of expert feedback in the context of large and small datasets, LTW1 and STW1 datasets are selected to inspect CAAD-EF performance. Table 6.2 showcases the experimental results. Incorporating expert feedback (i.e., CAAD-EF) yields a significant performance improvement in the case of large (LTW1) and small (STW1) training datasets. Specifically, incorporating expert feedback on a small subset of uncertain

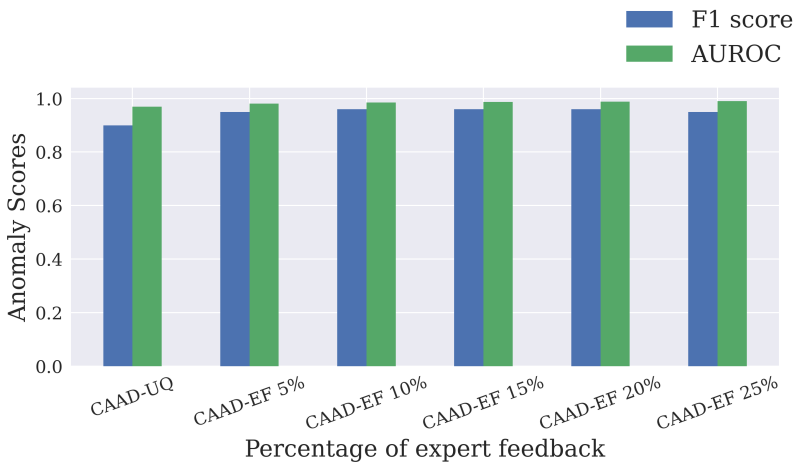

Figure 6.2: **Effect of expert feedback on model performance**. Improvement in Anomaly F1 scores and AUROC values are observed as the percentage of expert feedback us increased from 0% (CAAD-UQ) to 25% feedback. The model is able to show noticeable improvement in F1 score even with 5% expert feedback.

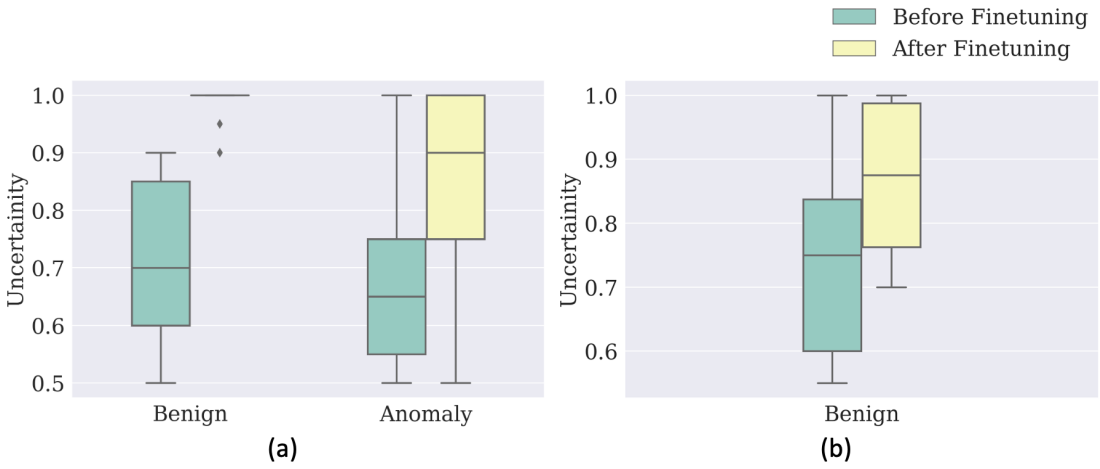

Figure 6.3: **Effect of expert feedback on uncertainty values**. These figures display box-plots of uncertainty values of human-in-the-loop (HIL) instances before and after finetuning. Figure (a) shows results from LTW1 dataset and figure (b) shows results from STW1 dataset. Figure (b) does not contain values for HIL anomalies as there were no HIL anomalies. Here, uncertainty value of 1 indicates high certainty of prediction and uncertainty value of 0 indicates low certainty of prediction.

instances (instances corresponding to 5% of the most uncertain test predictions as candidates for expert feedback) yields an average performance improvement of **4.17**% over the

next best model on the anomaly detection task (i.e., improvement in Anomaly F1 score) across both datasets. This indicates that the CAAD-EF benefits significantly from expert feedback in the context of different anomalies and data sizes. Additionally, CAAD-EF also showcases a **3.79**% performance improvement over CAAD (i.e., the variant without explicit feedback) in recognizing benign instances (i.e., improvement in Benign F1 score) across both the datasets showcasing a holistic performance improvement. These results are indicative of a highly effective, holistic and generic anomaly detection solution. For completeness, the effect of model performance of CAAD-EF with increase in expert feedback (Fig. 6.2) is further characterized. Figure 6.3 shows the values of uncertainty for the 5% most uncertain instances before finetuning (from CAAD-UQ) and after finetuning (from CAAD-EF). The improvement in uncertainty scores after finetuning can be clearly noticed. This result further clarifies the improved performance in CAAD-EF .

6.2.2 Effect of Expert Feedback (Qualitative Evaluation)

To further corroborate our claim of improved representation learning of CAAD-EF due to expert feedback, the evolution of the discriminator embeddings of CAAD-EF before and after fine-tuning with expert feedback are analysed. Fig 6.4 showcases t-SNE plots of embeddings inferred by the CAAD-EF discriminator. Fig. 6.4a shows the representations inferred by CAAD-UQ before expert feedback; color indicates ground truth labels (red: anomalies, blue: benign). The effect of the contrastive learning employed to train the discriminator, leads to clear separation of anomalous and benign regions in the plot. In Fig. 6.4b, it can be noticed that CAAD-UQ is uncertain about a significant number of points in the inference set. This region of uncertainty (ROU - indicated by dotted black circle in Fig. 6.4b) is identified and 5% most uncertain instances (as indicated by CAAD-UQ) are supplied to

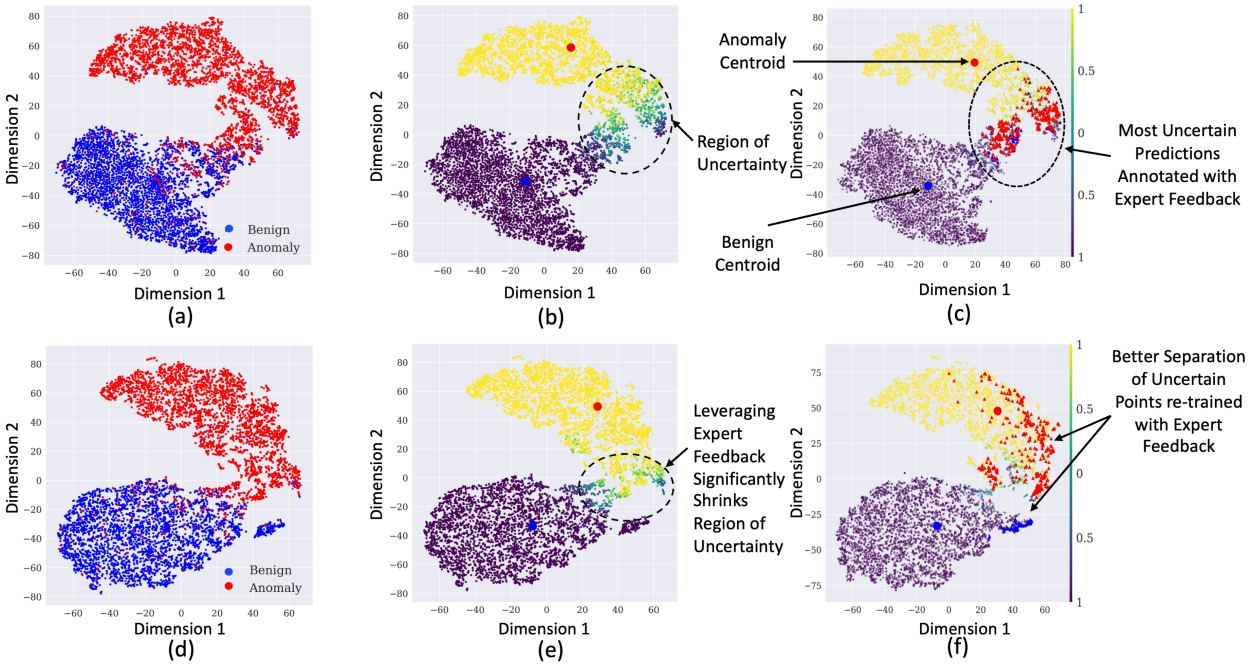

Figure 6.4: **Qualitative effect of expert feedback on model performance**. The figures (a) - (f) qualitatively represent the effect of incorporating human feedback in our proposed CAAD-EF framework on the LTW1 dataset. (a) Depicts t-SNE embeddings of our CAAD-UQ model discriminator, colored by the ground truth labels with anomalies colored red and benign points colored blue. (b) Showcases the same t-SNE embeddings as Fig. 6.4a but colored by the uncertainties obtained from the CAAD-UQ model (yellow: anomaly with low prediction uncertainty, purple: benign with low prediction uncertainty, green regions indicate uncertain instances;). Highly focused but sizeable 'Region of Uncertainty' (ROU) indicated by the dotted black circle can be noticed. (c) Ground truth labels of points in the ROU as specified by the expert (red: anomalous points, blue: benign points).(d) Depicts (similar to Fig. 6.4a) updated t-SNE embeddings yielded by the CAAD-UQ discriminator after fine-tuning with expert feedback for 5% most uncertain instances. (e) Updated uncertainty estimates of CAAD-UQ post fine-tuning, a significant reduction in ROU can be seen (indicated by dotted black circle) compared to Fig. 6.4b (f) 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. 6.4e.

the expert for feedback. These expert labeled instances are highlighted as red (expert label: anomaly) or blue (expert label: benign) points in Fig 6.4c. The model is fine-tuned with the full training set and the updated sets of points to produce new uncertainty estimates (Fig. 6.4e) wherein it can be seen that the model is significantly less uncertain in the ROU

(which has shrunk significantly). Finally, it can be noticed that the instances that were supplied by the expert as feedback have achieved significant separation and gravitated towards their respective cluster centroids (Fig. 6.4f) thereby leading to improved model performance in the CAAD-EF framework.

6.3 Ablation Study of CAAD-EF

Thus far, through rigorous qualitative and quantitative experiments, the effectiveness of our proposed CAAD-EF framework for anomaly detection has been verified. CAAD-EF consists of multiple facets and it is important to characterize the effect of each. Hence, a detailed ablation study of the proposed CAAD-EF framework is conducted. Table 6.3 details the results. From the table it can be noticed that the CAAD-EF model is dependent on each facet of its pipeline for effective representation learning with the most significant drop in performance occurring due to the removal of the contrastive learning based model training. It can also be noticed that the performance of CAAD is a function of the effect of contrastive learning and adversarial training. The performance is further improved with the inclusion of expert feedback (CAAD-EF). Removal of the UNet blocks from the generator also lead to deterioration in performance, primarily due to decrease in the generator learning capability. Finally, there is a drop in performance (0.94 to 0.91 Wt. Avg. F1) when expert feedback is ignored. The results in Table 6.3 further reinforce the effectiveness of CAAD-EF framework for the task of anomaly detection.

Table 6.1: **Summary of results**. This table demonstrates that our proposed CAAD model is effective across multiple types of anomalies in the communication network data while also being effective on other standard datasets like MNIST. CAAD also outperforms SOTA baselines significantly for the anomaly detection task (Anomaly F1. score) across all datasets.

Data	Model	Benign F1	Anomaly F1	Overall AUROC	Overall AUPRC	Overall Wt. Avg. 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 6.2: **Impact of fine tuning with expert feedback**. There is a significant improvement in performance for both benign and anomaly detection tasks after incorporating expert feedback (CAAD-EF) relative to the expert feedback agnostic models CAAD , CAAD-UQ . CAAD-EF 95% is the CAAD-EF evaluated only on a test-set with expert feedback instances removed.

Data	Model	Benign F1	Anomaly F1	Overall AUROC	Overall AUPRC	Overall Wt. Avg. 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

Table 6.3: **Incremental ablation of CAAD-EF** . The removal of important components of CAAD-EF incrementally, causes performance deterioration. EF: Finetuning after Expert Feedback, UQ: Uncertainty Quantification and CL: Contrastive Learning.

Model	Benign F1	Anomaly F1	Overall AUROC	Overall AUPRC	Overall Wt. Avg. F1
CAAD-EF	0.94	0.94	0.98	0.98	0.94
CAAD-EF w/o EF	0.92	0.9	0.97	0.98	0.91
CAAD-EF w/o EF, UQ	0.93	0.9	0.97	0.97	0.92
CAAD-EF w/o EF, UQ, CL	0.74	0.41	0.86	0.89	0.58
CAAD-EF w/o EF, UQ, CL, UNet	0.72	0.28	0.84	0.83	0.5
CAAD-EF w/o EF, UQ, CL, WGAN-GP	0.73	0.32	0.83	0.8	0.53

Chapter 7

Conclusions

A novel anomaly detection framework called CAAD , employing adversarial training and contrastive learning is introduced. Through rigorous experiments it is demonstrated that the proposed method outperforms SOTA anomaly detection baselines and achieves a **92.84**% improvement for anomaly detection in wireless communication networks as well as in more generic anomaly detection contexts. CAAD-EF , a variant of CAAD capable of incorporating expert feedback is introduced and its effectiveness is evaluated through several qualitative and quantitative experiments. Incorporating expert feedback gives a performance boost of 4.19% over CAAD . Finally, the importance of each facet of our proposed CAAD-EF framework is highlighted through a detailed ablation study.

Chapter 8

Future Directions

Although extensive experiments have been performed to evaluate the proposed model, more study is needed to improve the proposed framework.

- Moving forward CAAD-EF can be augmented with more sophisticated uncertainty quantification techniques and leverage the power of our model for real-time human-in-the-loop anomaly detection applications, especially plagued by covariate shift.
- The model can be evaluated on more varieties of sophisticated anomalies.
- More complicated datasets other than MNIST, such as image datasets can be used to evaluate generalisability.
- Metadata at the signal detection and extraction level can become more accurate, more efficient, and can be enriched with additional features.
- Additional feature spaces and feature representations exist, which could represent more of this information more efficiently for anomaly detection.
- This thesis uses only static datasets and evaluates the model on certain types of anomalies. However, in a real world scenario, data might continually evolve and so will the anomalies. Hence, continuous training and evaluation is crucial to evaluating the validity of the proposed model.

Bibliography

[1] Heba Abdelnasser, Moustafa Youssef, and Khaled A. Harras. Wigest: A ubiquitous wifi-based gesture recognition system. *2015 IEEE Conference on Computer Communications (INFOCOM)*, pages 1472–1480, 2015.

[2] Martin Arjovsky et al. Wasserstein gan, 2017.

[3] Ting Chen et al. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020.

[4] Haoqing Cheng, Heng Liu, Fei Gao, and Zhuo Chen. Adgan: A scalable gan-based architecture for image anomaly detection. In *ITNEC*, volume 1, pages 987–993. IEEE, 2020.

[5] Hyunsoo Cho, Jinseok Seol, and Sang-goo Lee. Masked contrastive learning for anomaly detection. *arXiv preprint arXiv:2105.08793*, 2021.

[6] Yeji Choi, Hyunki Lim, Heeseung Choi, and Ig-Jae Kim. Gan-based anomaly detection and localization of multivariate time series data for power plant. In *2020 IEEE international conference on big data and smart computing (BigComp)*, pages 71–74. IEEE, 2020.

[7] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR'05*, volume 1, pages 539–546. IEEE, 2005.

[8] Shubhomoy Das et al. Incorporating expert feedback into active anomaly discovery. In *ICDM*, pages 853–858. IEEE, 2016.

[9] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[10] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059. PMLR, 2016.

[11] Ian J. Goodfellow et al. Generative adversarial networks, 2014.

[12] Nico Görnitz et al. Toward supervised anomaly detection. *Journal of Artificial Intelligence Research*, 46:235–262, 2013.

[13] Ishaan Gulrajani et al. Improved training of wasserstein gans. *NeurIPS*, 30, 2017.

[14] Xu Han, Xiaohui Chen, and Li-Ping Liu. Gan ensemble for anomaly detection. *arXiv preprint arXiv:2012.07988*, 7(8), 2020.

[15] Wei Honghao, Jia Yunfeng, and Wang Lei. Spectrum anomalies autonomous detection in cognitive radio using hidden markov models. In *2015 IEEE Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pages 388–392. IEEE, 2015.

[16] Jongheon Jeong and Jinwoo Shin. Training gans with stronger augmentations via contrastive discriminator. *arXiv preprint arXiv:2103.09742*, 2021.

[17] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4037–4058, 2020.

[18] Prannay Khosla et al. Supervised contrastive learning. *NeurIPS*, 33:18661–18673, 2020.

[19] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 ICDM*, pages 413–422. IEEE, 2008.

[20] Song Liu, Yingying Chen, Wade Trappe, and Larry J Greenstein. Aldo: An anomaly detection framework for dynamic spectrum access networks. In *IEEE INFOCOM 2009*, pages 675–683. IEEE, 2009.

[21] Timothy J O'Shea, T Charles Clancy, and Robert W McGwier. Recurrent neural radio anomaly detection. *arXiv preprint arXiv:1611.00301*, 2016.

[22] Guansong Pang, Chunhua Shen, and Anton van den Hengel. Deep anomaly detection with deviation networks. In *SIGKDD*, pages 353–362, 2019.

[23] Guansong Pang, Anton van den Hengel, Chunhua Shen, and Longbing Cao. Toward deep supervised anomaly detection: Reinforcement learning from partially labeled anomaly data. In *ACM SIGKDD*, pages 1298–1308, 2021.

[24] Guansong Pang et al. Learning representations of ultrahigh-dimensional data for random distance-based outlier detection. In *SIGKDD*, pages 2041–2050, 2018.

[25] Sreeraj Rajendran, Wannes Meert, Vincent Lenders, and Sofie Pollin. Saife: Unsupervised wireless spectrum anomaly detection with interpretable features. In *2018 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, pages 1–9. IEEE, 2018.

[26] Sreeraj Rajendran, Wannes Meert, Vincent Lenders, and Sofie Pollin. Saife: Unsupervised wireless spectrum anomaly detection with interpretable features. In *2018 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, pages 1–9. IEEE, 2018.

[27] Sreeraj Rajendran, Vincent Lenders, Wannes Meert, and Sofie Pollin. Crowdsourced wireless spectrum anomaly detection. *IEEE Transactions on Cognitive Communications and Networking*, 6(2):694–703, 2019.

[28] Tal Reiss and Yedid Hoshen. Mean-shifted contrastive loss for anomaly detection. *arXiv preprint arXiv:2106.03844*, 2021.

[29] Olaf Ronneberger et al. U-net: Convolutional networks for biomedical image segmentation, 2015.

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

[31] Thomas Schlegl et al. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 54(January):30–44, 2019. ISSN 13618423. doi: 10.1016/j.media.2019.01.010.

[32] Nitish Srivastava et al. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15(1):1929–1958, 2014.

[33] Jihoon Tack et al. Csi: Novelty detection via contrastive learning on distributionally shifted instances, 2020.

[34] Nistha Tandiya, Ahmad Jauhar, Vuk Marojevic, and Jeffrey H Reed. Deep predictive coding neural network for rf anomaly detection in wireless networks. In *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*, pages 1–6. IEEE, 2018.

[35] Nistha Tandiya, Ahmad Jauhar, Vuk Marojevic, and Jeffrey H Reed. Deep predictive coding neural network for rf anomaly detection in wireless networks. In *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*, pages 1–6. IEEE, 2018.

[36] Hao Wang, Daqing Zhang, Yasha Wang, Junyi Ma, Yuxiang Wang, and Shengjie Li. Rt-fall: A real-time and contactless fall detection system with commodity wifi devices. *IEEE Transactions on Mobile Computing*, 16(2):511–526, 2017. doi: 10.1109/TMC.2016.2557795.

[37] Yanxin Wang, Johnny Wong, and Andrew Miner. Anomaly intrusion detection using one class svm. In *Proceedings from the Fifth Annual IEEE SMC Information Assurance Workshop, 2004.*, pages 358–364. IEEE, 2004.

[38] Dan Wu, Daqing Zhang, Chenren Xu, Yasha Wang, and Hao Wang. Widir: Walking direction estimation using wireless signals. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, page 351–362, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450344616. doi: 10.1145/2971648.2971658. URL https://doi.org/10.1145/2971648.2971658.

[39] Jure Zbontar et al. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, pages 12310–12320. PMLR, 2021.

[40] Xiaolong Zheng, Jiliang Wang, Longfei Shangguan, Zimu Zhou, and Yunhao Liu. Smokey: Ubiquitous smoking detection with commercial wifi infrastructures. In *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, page 1–9. IEEE Press, 2016. doi: 10.1109/INFOCOM.2016.7524399. URL https://doi.org/10.1109/INFOCOM.2016.7524399.

[41] Chong Zhou and Randy C Paffenroth. Anomaly detection with robust deep autoencoders. In *SIGKDD*, pages 665–674, 2017.

[42] Dali Zhu, Na Pang, Gang Li, and Shaowu Liu. Notifi: A ubiquitous wifi-based abnormal

activity detection system. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1766–1773, 2017. doi: 10.1109/IJCNN.2017.7966064.