Quantifying Raw RF Dataset Similarity for Transfer Learning Applications

LAUREN J. WONG1,2,3, SEAN MCPHERSON3, AND ALAN J. MICHAELS1,2 (Senior Member, IEEE)

1Hume Center for National Security and Technology, Virginia Tech, Blacksburg, VA 24060, USA
2Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA 24060, USA
3AI Lab, Intel Corporation, Santa Clara, CA 95054, USA

CORRESPONDING AUTHOR: L. J. WONG (e-mail: ljwong@vt.edu)

ABSTRACT Transfer learning (TL) has proven to be a transformative technology for computer vision (CV) and natural language processing (NLP) applications, offering improved generalization, state-of-the-art performance, and faster training time with less labelled data. As a result, TL has been identified as a key research area in the budding field of radio frequency machine learning (RFML), where deployed environments are constantly changing, data is hard to label, and applications are often safety-critical. TL literature and theory shows that TL is generally successful when the source and target domains and tasks are similar, but the term similar is not sufficiently defined. Therefore, quantifying dataset similarity is of importance for analyzing and potentially predicting TL performance, and also has further application in RFML dataset design. This work offers a dataset similarity metric, specifically designed for raw RF datasets, based on expert-defined features and χ² tests, and systematically evaluates the proposed metric using synthetic datasets with carefully curated signal-to-noise ratios (SNRs), frequency offsets (FOs), and modulation types. Results show that the proposed dataset similarity metric intuitively quantifies the notion of similar signal sets, so long as the expert-features used to construct the metric are well suited to the application.

INDEX TERMS Dataset similarity, deep learning (DL), machine learning (ML), radio frequency machine learning (RFML), transferability, transfer learning (TL).

I. INTRODUCTION

The application of machine learning (ML) and deep learning (DL) techniques in wireless communications settings has yielded state-of-the-art spectrum awareness, cognitive radio (CR), and networking algorithms. Such algorithms that utilize raw radio frequency (RF) data as input to ML/DL techniques are considered radio frequency machine learning (RFML) algorithms [1], [2]. Like all traditional ML techniques, most state-of-the-art RFML algorithms require copious amounts of labelled training data drawn from the intended deployment environment, and for the intended deployment environment to remain stable, in order to achieve said state-of-the-art performance [3]. Therefore, recent works have identified transfer learning (TL) as a key research thrust for RFML which would enable developers to train high-performing RFML models quickly and with less labelled data, compared to standard training practices, by using prior knowledge learned on a source domain/task for a target domain/task [3], [4]. Using TL techniques, state-of-the-art performance can be maintained across different deployment scenarios by tuning models to the intended platforms and channel environments. For example, a signal detection model trained to deploy on an aircraft carrier can be altered to deploy in a city center where the channel environment significantly differs, specific emitter identification (SEI) models can be modified to perform consistently across several unmanned aerial vehicles (UAVs) though the receiver hardware differs, and an automatic modulation classification (AMC) model can receive iterative updates to add and/or remove signals-of-interest as the threat landscape changes [3].

General TL theory specifies that TL is successful when the source/target domains and tasks are “similar” [5]. Therefore, TL performance should increase as the source and target...
domains and tasks become increasingly “similar.” However, the term “similar” is not sufficiently defined. In this work, we seek to quantify source/target domain and task similarity, through the use of a raw RF dataset similarity metric, as a means to better understand how and when TL is useful in the context of RFML. More specifically, given that the source and target datasets characterize the respective domains and tasks, source/target dataset similarity can be used as a proxy for source/target domain and task similarity.

In total, this work introduces a novel raw RF dataset similarity metric based on expert-defined features and $\chi^2$ tests, described in Sections III and IV-A and overviewed in Fig. 1. As in [6], the performance of the proposed metric is systematically evaluated using carefully curated synthetic datasets containing varying signal-to-noise ratios (SNRs), frequency offsets (FOs), and modulation schemes, for an example AMC use-case, described further in Section IV. The results of these experiments are presented in Sections V-A and V-B. Further, the correlation between the proposed dataset similarity metric and TL performance is examined in Section V-C, and the impact of dataset size and $\chi^2$ test parameters on the metric is discussed in Section V-D. Results show that the metric quantifies an intuitive notion of similarity, such that source/target dataset pairs increase in similarity through increasing overlap in modulation schemes, SNR ranges, and FO ranges, and that source/target dataset similarity and post-transfer accuracy are positively correlated. However, results also indicate that though dataset similarity, as quantified using the proposed similarity metric, and post-transfer accuracy are positively correlated, high similarity does not guarantee TL success. Therefore, we conclude the work in Section VI with a discussion on the utility of the metric with regards to analyzing and predicting transfer learning performance in research and deployed systems, and offer directions for improving the metric and other future work.

II. BACKGROUND & RELATED WORK

Despite abundant research in TL algorithms and performance in modalities such as computer vision (CV) and natural language processing (NLP) [7], very little work has examined TL algorithms and performance in RFML settings. As such, dataset similarity metrics such as information theoretic measures (i.e., Kullback-Leibler (KL) and Jensen-Shannon (JS) divergences) [8], higher-order measures (i.e., maximum mean discrepancy (MMD)) [9], principal component analysis (PCA)-based metrics [10], and the Proxy-A distance [11] have been considered for various CV, NLP, and multi-variate time series (MTS) applications.

Similar metrics have demonstrated success in feature-based CR settings, but have not yet demonstrated success with raw RF data which is high-dimensional, fast-changing, and highly dependent on the underlying bit pattern. For example, in [12], the KL divergence between historical measurement data such as received signal strength (RSS) and signal-to-interference-plus-noise ratio (SINR) is used to estimate the similarity between femtocells. Other works have made use of the Frechet Inception Distance (FID) evaluation index and MMD between neural network (NN) embeddings of spectograms [13], [14]. Several works have also examined the similarity between users in dynamic spectrum access (DSA) enabled CR networks which utilize between deep Q-networks. Such approaches quantify similarity between each pairs of secondary users in the network using the mean square error (MSE) between the action-value function parameters [15]. Finally, work in [16] uses dynamic time warping (DTW) to examine the similarity between frequency bands on an example-by-example basis for improving transfer learning performance in a cross-band spectrum prediction setting. This work differs from these existing approaches by:

- Comparing raw RF datasets which may differ in domain and/or task, rather than comparing measurement data history or action-value function parameters where the data is lower-dimensional and the task is generally held constant,
- Allowing the user to select appropriate RF-specific feature sets/weighting for the chosen use-case, rather than depending on NN embeddings, and
- Quantifying the similarity between entire raw RF datasets, as opposed to the similarity between individual examples.

This work overcomes these challenges by extracting a series of expert-defined features from each example in the source and target datasets, and comparing the distributions of these features using $\chi^2$ tests. The result is a flexible and computationally efficient dataset similarity metric, bounded between 0 and 1, that is applicable to both labelled and un-labelled datasets. By focusing solely on the datasets and
expert-defined features to estimate domain and task similarity, no ML/DL training/execution is required, minimizing compute, and the metric can be tailored to the application-of-interest (AMC, SEI, etc.). While there is work left to be done, discussed in length in Section VI, the experimental results show that dataset similarity presents a framework for model selection through dataset triage that is far more computationally efficient than blindly training NN models.

III. APPROACH

Because TL is generally understood to be successful when the source and target domains and/or tasks are “similar,” we propose a metric to measure the similarity between two raw in-phase/quadrature (IQ) datasets $d_1$ and $d_2$ through their expert-defined feature distributions, as shown in Fig. 1. By providing a quantifiable measure of similarity, this metric can be used to further examine when and how TL is productive, as was done with existing transferability metrics (Log Expected Empirical Prediction (LEEP) and Logarithm of Maximum Evidence (LogME)) in prior work for domain adaptation settings [6]. That is, we expect that the more similar the source and target domains and/or tasks, the better the transfer.

The metric is calculated as follows:

1. Calculate $n$ expert features for each sample in $d_1$ and $d_2$, creating feature vectors $f_1^0, ..., f_1^{n-1}$ and $f_2^0, ..., f_2^{n-1}$. The choice of expert features used in this step is dependent on the intended use-case, and is discussed further below.

2. For each feature vector, construct a histogram $(h_1^0, ..., h_1^{n-1}$ and $h_2^0, ..., h_2^{n-1})$ such that $h_i^0$ is the histogram corresponding to the feature vector $f_i^0$, $h_i^1$ is the histogram representation of $f_i^1$, and so on. Note that each histogram will have the same bin locations (start/stop).

3. Use $\chi^2$ tests to determine whether there is a statistically significant difference between $h_i^0$ and $h_i^0$, $h_i^1$ and $h_i^1$, ..., and $h_i^{n-1}$ and $h_i^{n-1}$, recording the p-values for each test ($p_0, ..., p_{n-1}$).

4. The final dataset similarity score is the weighted average of the p-values recorded:

$$\text{dataset similarity} = \frac{1}{n} \sum_{i=0}^{n-1} \alpha_i \cdot p_i$$

such that $\sum_{i=0}^{n-1} \alpha_i = 1$.

The following hyper-parameters and design options are to be chosen by the user, offering flexibility and customization:

- The choice of expert features used in step #1 is dependent upon the intended use-case, and directly impacts the usefulness of the metric in TL settings. For example, transient-based features might be used in an SEI setting [17], while higher-order statistics might be more specific to an AMC setting in which the datasets contain only PSK and QAM signals [18]. Therefore expert knowledge and/or feature selection methods should be used when customizing the proposed metric for a use-case-of-interest, a topic discussed further in Section VI.

- The number of bins per histogram. The impact of dataset size and the number of bins per feature histogram on the proposed dataset similarity score is further explored in Section V.

- The feature weightings $\alpha_0, ..., \alpha_n$.

Use-case and implementation details for the experiments and results presented herein are given in the following section.

Because $\chi^2$ p-values are bounded between 0 and 1, the proposed dataset similarity metric is also bounded between 0 and 1, such that 0 is most dissimilar and 1 is most similar. The metric is applicable to both labelled and unlabelled datasets, does not require the use of ML/DL models, and provides an intuitive notion of similarity, especially for traditional RF engineers who rely heavily on feature-based methods. Finally, the metric may also be extended to additional modalities (e.g., CV, NLP) through the use of expert-defined features specific to the respective modalities.

IV. EXPERIMENTAL SETUP

A. USE-CASE & IMPLEMENTATION

This work considers an example AMC application over the 23 signal types given in Table 1 which have been used in several prior works [6], [19]. The metric is calculated using the 7 time domain features given in Table 2. These features represent a baseline set of features used in existing feature-based AMC works [18], [20], and do not represent an “optimal” feature set for the given use-case. Rather, this work provides a proof-of-concept raw RF dataset similarity metric, despite a non-perfect set of features. In practice, we will not know the “optimal” set of features for a new problem, given real-world uncertainties, but augmenting the proposed metric with feature selection methods a promising path for future work and is discussed in greater detail in Section VI.

For simplicity, all features are weighted equally, such that

$$\alpha_0 = \cdots = \alpha_{n-1} = \frac{1}{n}.$$

Finally, dataset similarity is computed using 500 examples per output class, and the number of bins per feature histogram is set as

$$n_{\text{bins}} = \left\lceil \sqrt{500 \cdot m} \right\rceil$$

where $m$ is the number of output classes in the smaller of the two input datasets.

B. DATASETS

Using the master dataset created in [6] and publicly available on IEEE DataPort [21], data-subsets with carefully selected metadata parameters are constructed from the larger master dataset to evaluate the impact of SNR, FO, and modulation scheme, on the proposed dataset similarity metric. The master dataset contains 600000 examples of each of the
TABLE 1. Signal types included in this work.

| Modulation Name | Linear | Freq-shifted | Analog | All | Small | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 | S9 | S10 |
|-----------------|--------|--------------|--------|-----|------|----|----|----|----|----|----|----|----|----|-----|
| BPSK            | ✔      | ✔            | ✔      | ✔   | ✔    | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔   | ✔   |
| QPSK            | ✔      | ✔            | ✔      | ✔   | ✔    | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔   | ✔   |
| PSK8            | ✔      | ✔            | ✔      | ✔   | ✔    | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔   | ✔   |
| PSK16           | ✔      | ✔            | ✔      | ✔   | ✔    | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔   | ✔   |
| QAM64           | ✔      | ✔            | ✔      | ✔   | ✔    | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔   | ✔   |
| APSK16          | ✔      | ✔            | ✔      | ✔   | ✔    | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔   | ✔   |
| FSK5k           | ✔      | ✔            | ✔      | ✔   | ✔    | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔   | ✔   |
| FSK5k           | ✔      | ✔            | ✔      | ✔   | ✔    | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔   | ✔   |
| GFSK5k          | ✔      | ✔            | ✔      | ✔   | ✔    | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔   | ✔   |
| GFSK75k         | ✔      | ✔            | ✔      | ✔   | ✔    | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔   | ✔   |
| MSK             | ✔      | ✔            | ✔      | ✔   | ✔    | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔   | ✔   |
| GMSK            | ✔      | ✔            | ✔      | ✔   | ✔    | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔   | ✔   |
| FM-NB           | ✔      | ✔            | ✔      | ✔   | ✔    | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔   | ✔   |
| FM-WB           | ✔      | ✔            | ✔      | ✔   | ✔    | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔   | ✔   |
| AM-DSB          | ✔      | ✔            | ✔      | ✔   | ✔    | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔   | ✔   |
| AM-DSBSC        | ✔      | ✔            | ✔      | ✔   | ✔    | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔   | ✔   |
| AM-LSB          | ✔      | ✔            | ✔      | ✔   | ✔    | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔   | ✔   |
| AM-USB          | ✔      | ✔            | ✔      | ✔   | ✔    | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔   | ✔   |
| AWGN            | ✔      | ✔            | ✔      | ✔   | ✔    | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔  | ✔   | ✔   |

signal types given in Table 1, synthesized in an additive white Gaussian noise (AWGN) environment, with the SNR and FO of a given example selected uniformly at random between $[-10\text{dB}, 20\text{dB}]$ and $[-10\%, 10\%]$ of the sample rate, respectively, and no phase offset. All additional signal generation parameters can be found in [6]. The data-subsets were constructed from the larger master dataset as follows:

To evaluate the impact of SNR and FO on transfer learning performance and transferability, three sweeps were performed across SNR, FO, and SNR + FO. For the SNR sweep, 26 source data-subsets were constructed from the master dataset, each containing examples with SNRs selected uniformly at random from a 5dB range sweeping from $[-10\text{dB}, 20\text{dB}]$ in 1dB steps (i.e., $[-10\text{dB}, -9\text{dB}], [-9\text{dB}, -8\text{dB}], \ldots, [15\text{dB}, 20\text{dB}]$), and FO selected uniformly at random between $[-5\%, 5\%]$ of sample rate. For the FO sweep, 31 source data-subsets were constructed from the master dataset, each containing examples with FOs selected uniformly at random from a 5% range sweeping from $-10\%$ of sample rate to $10\%$ of sample rate in 0.5% steps (i.e., $[-10\%, -7.5\%], [-7.5\%, -5\%], \ldots, [0\%, 10\%]$), and SNR selected uniformly at random between $[0\text{dB}, 20\text{dB}]$. For the SNR + FO sweep, 25 source data-subsets were constructed from the master dataset containing examples with SNRs selected uniformly at random from a 10dB range sweeping from $-10\text{dB}$ to $20\text{dB}$ in 5dB steps (i.e., $[-10\text{dB}, 0\text{dB}], [-5\text{dB}, 5\text{dB}], \ldots, [10\text{dB}, 20\text{dB}]$) and with FOs selected uniformly at random from a 10% range sweeping from $-10\%$ of sample rate to $10\%$ of sample rate in 2.5% steps (i.e., $[-10\%, 0\%], [-7.5\%, 2.5\%], \ldots, [0\%, 10\%]$).

To investigate transfer across broad categories of modulation types, namely linear, frequency-shifted, and analog modulation schemes, 5 source data-subsets were constructed from the larger master dataset containing the following modulation schemes:

- All modulations
- Small Subset = BPSK, QPSK, OQPSK, QAM16, QAM64, APSK5k, FSK 5k, MSK, FM-NB, DSB, USB, AWGN
- Linear modulations = BPSK, QPSK, PSK8, PSK16, OQPSK, QAM16, QAM32, QAM64, APSK16, APSK32, AWGN
- Frequency-shifted modulations = FSK 5k, FSK 75k, GFSK 5k, GFSK 75k, MSK, GMSK, AWGN
- Analog modulations = FM-NB, FM-WB, DSB, DBSC, LSB, USB, AWGN

For each data-subset in this modulation type experiment, SNR was selected uniformly at random between $[0\text{dB}, 20\text{dB}]$ and FO was selected uniformly at random between $[-5\%, 5\%]$ of sample rate. This experiment is denoted “Modulation Types” in all figs. and tables.

A second set of data-subsets was constructed such that a single modulation type was added/removed from the small/all...
TABLE 2. The instantaneous time domain features used in this work, where $a(t)$, $\phi(t)$, and $f_N(t)$ are the instantaneous amplitude, phase, and frequency of the example $s(t)$, respectively.

| Feature | Formula |
|---------|---------|
| The standard deviation of the absolute value of the normalized instantaneous amplitude of the example | $\sigma_a = \left[ \frac{1}{N_s} \sum_{i=1}^{N_s} a_n^2(i) \right]^{1/2}$ |
| The standard deviation of the instantaneous phase of the example | $\sigma_p = \left[ \frac{1}{N_s} \sum_{i=1}^{N_s} \phi^2(i) \right]^{1/2}$ |
| The standard deviation of the absolute value of the instantaneous phase of the example | $\sigma_{ap} = \left[ \frac{1}{N_s} \sum_{i=1}^{N_s} |\phi(i)| \right]^{1/2}$ |
| The standard deviation of the absolute value of the normalized instantaneous frequency of the example | $\sigma_f = \left[ \frac{1}{N_s} \sum_{i=1}^{N_s} f_n^2(i) \right]^{1/2}$ |
| The kurtosis of the instantaneous amplitude | $K_a = \frac{\sum_{i=1}^{N_s} a(i) - m_a)^4}{(\sum_{i=1}^{N_s} a(i) - m_a)^2}$ |
| The kurtosis of the instantaneous frequency | $K_f = \frac{\sum_{i=1}^{N_s} f(i) - m_f)^4}{(\sum_{i=1}^{N_s} f(i) - m_f)^2}$ |
| Spectrum symmetry | $P = \frac{P_L - P_U}{P_L + P_U}$ |

where $P_L = \sum_{i=1}^{f_0} S(i)^2$, $P_U = \sum_{i=1}^{f_s} S(i) - f_0 + 1)^2$, $S(t)$ is the frequency domain representation of $s(t)$, and $f_0$ is the sample number corresponding to the carrier frequency.

modulations datasets described above, mimicking a successive model refinement scenario. This experiment is denoted “Model Refinement” in all figs. and tables. More specifically, the 12 source data-subsets were constructed from the larger master dataset containing:

- Small = BPSK, QPSK, OQPSK, QAM16, QAM64, APSK16, FSK 5k, MSK, FM-NB, DSB, USB, AWGN
- S1 = Small + PSK8
- S2 = S1 + PSK16
- S3 = S2 + QAM32
- S4 = S3 + APSK32
- S5 = S4 + FSK 75k
- S6 = S5 + GFSK 5k
- S7 = S6 + GFSK 75k
- S8 = S7 + GMSK
- S9 = S8 + FM-WB
- S10 = S9 + DSBSC
- All = S10 + LSB

Again, SNR was selected uniformly at random between [0dB, 20dB] and FO was selected uniformly at random between [−5%, 5%] of sample rate.

C. MODEL ARCHITECTURE & TRAINING

All models trained for the purposes of evaluating the proposed dataset similarity metric use the same convolutional neural network (CNN) architecture used in the prior work [6] and shown in Table 3. Other model architectures tested showed the same trends, but are not included in this work for brevity.

Standard model pre-training and TL pipelines and hyper-parameter settings, like those used previously, are also used herein and shown in Fig. 2. More specifically, all source models utilized the “full”-sized data-subsets, which contained 5000 and 500 training and validation examples per class respectively. Further, all source models were trained using the Adam optimizer [22], Cross Entropy Loss [23], a learning rate of 0.001, and without weight decay for a total of 100 epochs, with checkpoints saved when the lowest validation loss was achieved and reloaded at the conclusion of training.

TL was performed using the “limited”-size data-subsets which contained 500 and 50 training and validation examples per class respectively. All results shown herein utilize head re-training, as prior work showed that head re-training to be as
FIGURE 2. The model pre-training and TL pipeline.

| Layer Type | Num Kernels/Nodes | Kernel Size |
|------------|-------------------|-------------|
| Input      | size = (2, 128)   |             |
| Conv2d     | 1500              | (1, 7)      |
| ReLU       | 96                | (2, 7)      |
| Dropout    | rate = 0.5        |             |
| Flatten    |                   |             |
| Linear     | 65                |             |
| Linear     | m                 |             |

Trainable Parameters: 7432725 + (66 \cdot m)

TABLE 3. Model architecture with \( m \) being the number of output classes (modulation types) trained.

effective, if not more effective, than fine-tuning for RF TL performance [24]. For head re-training, the final layer of the model was trained using a learning rate of 0.001 and without weight decay, while the rest of the model’s parameters were frozen. Head-retraining was also performed using the Adam optimizer, Cross Entropy Loss, and with checkpoints saved at the lowest validation loss over 100 epochs.

Finally, a set of baseline models were trained on both the “full”-size and “limited”-size data-subsets. These baseline models were from random initialization using the same training hyper-parameters described for pre-training (the Adam optimizer, Cross Entropy Loss, etc.).

V. RESULTS & ANALYSIS

First, in Sections V-A and V-B, we show that the proposed metric is consistent with an intuitive understanding of dataset similarity, expecting similarity to increase as the intersection or overlap in SNR, FO, and modulation schemes in the source/target datasets increases. Then, in Section V-C we verify that similar source/target datasets, as quantified by the proposed metric, result in successful RF TL by showing that the proposed metric positively correlates with improved performance of the TL models over baseline models. Finally, Section V-D examines how dataset size and the \( \chi^2 \) test parameters impact the proposed dataset similarity metric.

A. IMPACT OF SNR AND FO ON DATASET SIMILARITY

As previously described, the same data-subsets constructed to evaluate the impact of SNR and FO on transfer learning performance and transferability in [6] were used in this work to evaluate the proposed dataset similarity metric. Figs. 3–5 show the proposed similarity metric across the data-subsets with varying SNR, FO, and SNR + FO respectively, with all other metadata parameters held constant. As expected, for each of these parameter sweeps, similarity increases the more intuitively similar the parameter range (i.e., \([-5dB, 0dB]\) is closer to \([-4dB, 1dB]\) than \([0dB, 5dB]\)). However, Figs. 3 and 5 expose the sensitivity of the metric to relatively small changes in SNR, a result of the expert features used herein which are also sensitive to noise. This effect could likely be mitigated using a different selection of expert features, a topic discussed further in Section VI. Echoing the results given in [6], Figs. 3–5
also show SNR has a much larger impact on similarity than FO. Also of note, Fig. 4 shows that similarity, as quantified by the proposed metric is symmetric, with regards to FO. That is, datasets with equal amounts of FO in the positive and negative direction (i.e., [−10%, −5%] of sample rate and [5%, 10%] of sample rate) are highly similar. While similarity along the primary diagonal is intuitive, similarity along the secondary diagonal is again a result of the expert features used in this work, as constant positive and negative rotations throughout the duration of the example capture result in the same instantaneous time domain features.

B. SIMILARITY OF DATASETS CONTAINING DIFFERENT MODULATION SCHEMES

Fig. 6 shows the similarity across broad categories of modulation types, such that the linear, frequency-shifted, and analog modulation type datasets share no modulation schemes other than of the AWGN noise class. As expected, similarity between datasets with no shared modulation schemes is very low. In contrast, in Fig. 7, Small is a subset of Subset1 which is a subset of Subset2 and so on. As a result, similarity is highest along the diagonal, where the data-subsets share the most modulation schemes, as expected.

C. CORRELATIONS BETWEEN SIMILARITY AND TL PERFORMANCE

As previously discussed, the proposed metric builds upon the general understanding that TL is successful when source/target domains are “similar.” Therefore, in this section, the accuracy of TL models is examined as a function of
the proposed metric, and compared to the baseline models described in Section IV-C.

Figure 8 shows the difference between the accuracy of TL models versus baseline models as a function of dataset similarity for the SNR and FO sweeps described above, as well as the modulation type experiments. Additionally, Table 4 shows the Pearson’s $r$ correlation coefficient between difference between the accuracy of TL models versus baseline models and the proposed dataset similarity metric.

| Dataset Similarity Metric | Full Data Baseline | Limited Data Baseline |
|---------------------------|--------------------|-----------------------|
| SNR Sweep                 | 0.5634             | 0.5799                |
| FO Sweep                  | 0.5087             | 0.4333                |
| SNR + FO Sweep            | 0.5243             | 0.5364                |
| Modulation Types          | 0.2440             | 0.2860                |
| Model Refinement          | 0.2293             | 0.2494                |

Given that a high/low similarity score does not directly infer TL performance, the question becomes: “Is this metric useful in helping to determine whether it is more beneficial to use TL approaches or train from random initialization?” It should first be noted that causality is not expected, as dataset similarity is known to only be one facet of TL performance. Rather, the proposed dataset similarity is a tool for rank prioritizing the raw RF datasets most likely to be fruitful for TL in an RFML setting. However, in the case where the size of the target dataset is limited (as shown in Fig. 8(a), for some use-cases there is a similarity threshold at which all models perform better using TL over training from random initialization. For example, Fig. 9 isolates the points in 8(a) representing the Model Refinement scenario, and shows that such a threshold is present when similarity is greater than 0.63.

D. IMPACT OF DATASET SIZE AND HISTOGRAM BIN SIZE ON SIMILARITY

Throughout each of the experiments mentioned above, the proposed dataset similarity metric was computed using 500 examples per output class and setting $n_{\text{bins}} = \lfloor \sqrt{500 \cdot m} \rfloor$. Because $n$ features are computed for each example in each dataset, the proposed dataset similarity metric operates on histograms of those features, and the number of histogram bins is selected as a function of dataset size, we expect that the metric is largely invariant to dataset size, so long as a sufficient number of examples are available to characterize the dataset. This “sufficient” number of examples is likely subject to the diversity of the datasets, with more
FIGURE 9. The difference between the accuracy of TL models versus baseline models, in decimal format, as a function of the proposed metric for the Model Refinement use-case. When the difference is above zero, as shown by the solid black line, the TL model outperforms the baseline model.

FIGURE 10. Dataset similarity as a function of dataset size and the resultant number of bins per histogram used in each $\chi^2$ test for the first column of source/target dataset pairs shown in Fig. 7, averaged over 10 iterations. Results show that when the datasets are too small (i.e., 100 examples per dataset), similarity is artificially high for all dataset pairs. As dataset size increases, similarity steadily decreases. However, similarity is relatively consistent between source/target pairs with as few as 900 examples per dataset and 30 bins per histogram, meaning the similarity between the Small and S2 data-subsets. While larger datasets with more bins per histogram are generally advantageous when computing the proposed similarity metric, as the separation between dataset pairs increases, meaningful conclusions can be drawn from much smaller datasets.

VI. CONCLUSION & FUTURE WORK
Motivated by the notion that in TL settings source/target dataset similarity is correlated with post-transfer performance, this work has presented a novel RF dataset similarity metric based on expert-defined features and $\chi^2$ tests. The metric is bounded between 0 and 1, and does not rely on computationally-expensive ML/DL training or evaluation methods. Results have shown that the proposed metric quantifies dataset similarity intuitively, with similarity increasing the more similar the parameter ranges and modulation types within the datasets. However, results have also shown that while dataset similarity and TL performance are positively correlated, high dataset similarity does not ensure TL success or vice versa. This section outlines the numerous directions for future research pertaining to feature selection methods and criteria in particular, as well as ways that raw RF dataset similarity could be used to enhance current state-of-the-art RFML training methods for increased performance and reduced computation costs.

The instantaneous time domain features used in this work provide baseline results for an AMC setting, they do not represent an “optimal” feature set, and would likely not be sufficient in an SEI setting to identify similarities/differences between emitters. The results presented herein have also shown that the proposed dataset similarity metric can be sensitive to changes in SNR; a direct result of the same sensitivity in the instantaneous time domain features used to construct the metric. In an AMC setting alternative features such as wavelet-based features, cyclic-cumulant-based features, and higher order statistics would likely offer better robustness to noise, at the cost of increased computational complexity [20].

In order for the proposed dataset similarity metric to be most useful across different RF use-cases, expert-defined feature selection is of the utmost importance, and identifying better feature selection criteria and/or combining the proposed dataset similarity metric with formal feature selection methods would increase the utility of the proposed metric across tasks and domains. It should be noted that the problem of identifying the “optimal” features for quantifying raw RF dataset similarity is ill-defined. That is, the notion of RF dataset similarity likely changes depending on the use-case-of-interest, and no ground-truth similarity score exists. However, at minimum, selecting better feature sets for quantifying raw RF dataset similarity should consider both the relevance of the features to the use-case-of-interest, as well as maximizing the orthogonality between features to minimize redundancy. Future feature selection methods could tackle these criteria separately or in tandem.
Starting with the concept of feature relevance, two approaches can be taken: using expert knowledge to identify relevant feature from which to perform feature selection, or creating some notion of “ground truth” similarity (using post-transfer accuracy, for example) that can be used to selecting features relevant from a wide pool of candidate features. When considering feature orthogonality, possible approaches include computing the correlation coefficient between candidate features for a single dataset [26], using distance metrics such as the KL-divergence [27], JS-divergence [28], or Bhattacharyya distance [29] to select for dissimilarity between feature distributions for individual datasets and/or between datasets. If considering feature relevance and orthogonality in tandem and assuming some user-defined notion of “ground truth” similarity, possible feature selection methods also include Principal Feature Analysis (PFA) [30] and genetic algorithms [31].

The results in Section V show the positive correlation between the proposed raw RF dataset similarity metric and TL performance. This relationship indicates that dataset similarity may be a sufficient measure to suggest whether or not a source dataset is relevant to a target domain/task, with some degree of inherent uncertainty associated with correlation. These results do not show causation or suggest that a high dataset similarity score results in improved TL performance (or vice versa). Were the proposed dataset similarity metric and TL performance to have a causal relationship, it would be feasible to reason about the learned behaviors of NNs trained on raw RF datasets using the proposed metric.

Assuming appropriate expert-defined features have been selected for the intended use-case, the ability to intuitively quantify similarity between RF datasets is useful in both research and deployment settings. For example, a dataset similarity metric helps bolster the type of analysis performed in prior work [6], where RF TL behavior may be better understood when viewed through the lens of dataset similarity. To extend this work, the proposed dataset similarity metric could be combined with existing transferability metrics, such as LEEP and LogME, to more robustly predict TL performance more by considering the data directly, rather than just a ML/DL model’s response to the data. The proposed dataset similarity metric also provides a method to detect changes in the RF environment, which may necessitate re-training of deployed RFML models. Further, results presented in Section V-C showed that dataset similarity thresholds can be identified that indicate a TL approach will outperform training from random initialization, especially for data-limited settings. Finally, the proposed metric could be used to design more robust RFML datasets by identifying thresholds that indicate benefits from combining data-subsets into a larger training set based on the similarity of the data-subsets. For example, in an AMC setting, as combining less similar data-subsets that contain the same modulation schemes is likely to increase generalization performance.

REFERENCES

[1] T. Rpondeau. “Radio frequency machine learning systems (RFMLs).” Accessed: Jun. 2022. [Online]. Available: https://www.darpa.mil/program/radio-frequency-machine-learning-systems
[2] L. J. Wong, W. H. Clark, B. Flowers, R. M. Buehrer, W. C. Headley, and A. J. Michaels, “An RFML ecosystem: Considerations for the application of deep learning to spectrum situational awareness,” IEEE J. Commun. Netw., vol. 2, pp. 2243–2264, 2020.
[3] L. J. Wong and A. J. Michaels, “Transfer learning for radio frequency machine learning: A taxonomy and survey,” Sensors, vol. 22, no. 4, p. 1416, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/4/1416
[4] S. Kuzdeba, J. Robinson, and J. Carmack, “Transfer learning with radio frequency signals,” in Proc. IEEE 18th Annu. Consum. Commun. Conf. (ICCC), 2020.
[5] S. J. Pan and Q. Yang, “A survey on transfer learning,” IEEE Trans. Knowl. Data Eng., vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
[6] L. J. Wong, S. McPherson, and A. J. Michaels, “Assessing the value of transfer learning metrics for RF domain adaptation,” 2022, arXiv:2206.08329.
[7] S. Ruder, “Neural transfer learning for natural language processing,” Ph.D. dissertation, Neural Transf. Learn. Nat. Lang. Proc., Nat. Univ. Ireland, Galway, Ireland, 2019.
[8] B. Plank and G. van Noord, “Effective measures of domain similarity for parsing,” in Proc. 49th Annu. Meeting Assoc. Comput. Linguist. Hum. Lang. Technol. Jun. 2011, pp. 1566–1576. [Online]. Available: https://aclanthology.org/P11-1157
[9] W. Zellinger, T. Grubinger, E. Luhgofer, T. Natschläger, and S. Saminger-Platz, “Central moment discrepancy (CMD) for domain-invariant representation learning,” 2017, arXiv:1702.08831.
[10] K. Yang and C. Shahabi, “A PCA-based similarity measure for multivariate time series,” in Proc. 2nd ACM Int. Workshop Multimedia Databases, 2004, pp. 65–74.
[11] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, “Analysis of representations for domain adaptation,” in Advances in Neural Information Processing Systems, vol. 19, B. Schölkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA, USA: MIT Press, 2006. [Online]. Available: http://www.john.blitzer.com/papers/nips06.pdf
[12] W. Wang, J. Zhang, and Q. Zhang, “Transfer learning based diagnosis for configuration troubleshooting in self-organizing femtocell networks,” in Proc. IEEE Global Telecommun. Conf., 2011, pp. 1–5.
[13] W. H. Clark et al., “Developing RFML intuition: An automatic modulation classification architecture case study,” in Proc. IEEE Mil. Comm. Conf. (MICOM), 2019, pp. 292–298.
[14] O. A. Dobre, A. Abd, Y. Bar-Ness, and W. Su, “Survey of automatic modulation classification techniques: Classical approaches and new trends,” IET Commun., vol. 1, no. 2, pp. 137–156, 2007.
[15] L. J. Wong, S. McPherson, and A. J. Michaels, “Transfer learning for RF domain adaptation—Synthetic dataset,” Mar. 2022. [Online]. Available: https://ieee-dataport.org/open-access/transfer-learning-rf-domain-adaptation—synthetic-dataset
[16] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980.
[23] “Cross entropy loss.” Accessed: Jun. 2022. [Online]. Available: https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html

[24] L. J. Wong, S. McPherson, and A. J. Michaels, “An analysis of RF transfer learning behavior using synthetic data,” 2022, arXiv:2210.01158.

[25] P. Schober, C. Boer, and L. A. Schwarte, “Correlation coefficients: Appropriate use and interpretation,” Anesthesia Analgesia, vol. 126, no. 5, pp. 1763–1768, 2018.

[26] J. Benesty, J. Chen, Y. Huang, and I. Cohen, “Pearson correlation coefficient,” in Noise Reduction in Speech Processing. Cham, Switzerland: Springer, 2009, pp. 1–4.

[27] S. Kullback and R. A. Leibler, “On information and sufficiency,” Ann. Math. Stat., vol. 22, no. 1, pp. 79–86, 1951.

[28] J. Lin, “Divergence measures based on the Shannon entropy,” IEEE Trans. Inf. Theory, vol. 37, no. 1, pp. 145–151, Jan. 1991.

[29] A. Bhattacharyya, “On a measure of divergence between two multinomial populations,” Sankhyā Indian J. Stat., vol. 7, no. 4, pp. 401–406, 1946.

[30] Y. Lu, I. Cohen, X. S. Zhou, and Q. Tian, “Feature selection using principal feature analysis,” in Proc. 15th ACM Int. Conf. Multimedia, 2007, pp. 301–304.

[31] R. Leardi, R. Boggia, and M. Terrile, “Genetic algorithms as a strategy for feature selection,” J. Chemometr., vol. 6, no. 5, pp. 267–281, 1992.

LAUREN J. WONG received the B.A. degree in computer science and mathematics from Oberlin College, Oberlin, OH, USA, in 2016, the M.S. degree in electrical engineering from Virginia Tech, Blacksburg, VA, USA, in 2018, where she is currently pursuing the Ph.D. degree.

From 2018 to 2020, she was a Research Associate with the Hume Center for National Security and Technology, Virginia Tech. Since 2020, she has been a Research Scientist with Intel Corporation in the AI Lab. Her research interests include various aspects of RFML, from applications to operational considerations and user assurance, as well as data efficient learning and AI robustness.

SEAN MCPHERSON received the B.S. degree in electrical engineering from the University of Miami, Coral Gables, FL, USA, in 2004, and the M.S. and Ph.D. degrees in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 2007 and 2011, respectively.

From 2011 to 2018, he was a Systems Engineer with Northrop Grumman. Since 2018, he has been a Research Scientist with Intel Corporation in the AI Lab. His research interests include applications of signal processing to computer network traffic analysis, machine learning for cybersecurity, and data efficient machine learning.

ALAN J. MICHAELS (Senior Member, IEEE) received the degree in electrical and computer engineering, applied maths, and operations research from Georgia Tech and the MBA degree from Carnegie Mellon.

He serves as a Professor with the Virginia Tech’s Bradley Department of Electrical and Computer Engineering and the Director of the National Security Institute’s Spectrum Dominance division, Blacksburg, VA, USA. Previously, he worked a decade as a Systems Engineer, a Researcher, and the Department Head with Harris Corporation, specializing in secure communication. He is a Professional Engineer in the Commonwealth of Virginia and has 44 issued U.S. patents.

Dr. Michaels is a Fellow of the National Academy of Inventors.