Dimensional Reduction Analysis for Physical Layer Device Fingerprints with Application to ZigBee and Z-Wave Devices

Trevor J. Bihl and Kenneth W. Bauer Jr.
Department of Operational Sciences
Air Force Institute of Technology
Wright Patterson AFB, OH 45433
{Trevor.Bihl; Kenneth.Bauer}@afit.edu

Michael A. Temple and Benjamin Ramsey
Department of Electrical & Computer Engineering
Air Force Institute of Technology
Wright Patterson AFB, OH 45433
{Michael.Temple; Benjamin.Ramsey}@afit.edu

***Abstract*—Radio Frequency RF Distinct Native Attribute (RF-DNA) Fingerprinting is a PHY-based security method that enhances device identification (ID). ZigBee 802.15.4 security is of interest here given its widespread deployment in Critical Infrastructure (CI) applications. RF-DNA features can be numerous, correlated, and noisy. Feature Dimensional Reduction Analysis (DRA) is considered here with a goal of: 1) selecting appropriate features (feature selection) and 2) selecting the appropriate number of features (dimensionality assessment). Five selection methods are considered based on Generalized Relevance Learning Vector Quantization-Improved (GRLVQI) feature relevance ranking, and *p*-value and test statistic rankings from both the two-sample Kolmogorov-Smirnov (KS) Test and the one-way Analysis of Variance (ANOVA) F-test. Dimensionality assessment is considered using previous qualitative (subjective) methods and quantitative methods developed herein using data covariance matrices and the KS and F-test *p*-values. ZigBee discrimination (classification and ID verification) is evaluated under varying signal-to-noise ratio (*SNR*) conditions for both authorized and unauthorized rogue devices. Test statistic approaches emerge as superior to *p*-value approaches and offer both higher resolution in selecting features and generally better device discrimination. With appropriate feature selection, using only 16% of the data is shown to achieve better classification performance than when using all of the data. Preliminary first-look results for Z-Wave devices are also presented and shown to be consistent with ZigBee device fingerprinting performance.**

Keywords—ANOVA, dimensionality reduction, GRLVQI, F-test, feature selection, Kolmogorov-Smirnov, network security, physical layer, RF-DNA, ZigBee, Z-Wave

I. INTRODUCTION

Wireless Personal Area Network (WPAN) technologies, such as ZigBee and Z-Wave, enable low-power, low-cost mesh networks of numerous smart devices. WPANs are in active use throughout military and commercial enterprises, from hospitals [1, 2] to industrial control systems and transportation monitoring [3, 4]. ZigBee devices can form networks containing up to 65,000 devices while Z-Wave networks can include up to 232 devices [5]. Robust WPAN security is essential because they connect more devices to the physical world than any other wireless technology [6].

Low-cost wireless embedded systems do not possess the computational power or physical security of traditional computing devices. As a result, WPAN defense must be implemented within all OSI stack layers [7]. One of the most novel and robust techniques for identifying suspicious wireless activity is Physical Layer (PHY) fingerprinting. Encryption key-based measures generally neglect useful PHY information as an element of multi-factor authentication that includes [3]:

1. "Something you know" (NWK – encryption keys)
2. "Something you have" (MAC – MAC address)
3. "Something you are" (PHY – RF Fingerprints).

PHY layer characteristics are unique to each device and result from production variances and operational conditions, and therefore are considered as an additional, more robust, level of security. Additional reasons also exist for examining the PHY layer, including: authentication, intrusion detection, malfunction detection, and rogue access [8].

RF Distinct Native Attribute (RF-DNA) Fingerprinting uses statistical methods of feature extraction, classification (one vs. many), and ID verification (one vs. one) for device discrimination [9]. RF-DNA enables both classification and verification and has shown practical utility for both cross-model [10], e.g. similar devices from different manufacturers, and like-model (serial number) device discrimination [11].

In operational environments, RF-DNA fingerprint features can be numerous and of varying saliency. Therefore, this work addresses Dimensional Reduction Analysis (DRA) using ZigBee and Z-Wave devices as the application of interest. This work extends previous DRA studies [3, 12, 13] by examining and comparing results from five DRA methods: Generalized Relevance Learning Vector Quantization-Improved (GRLVQI) feature relevance ranking, and *p*-values and test statistic values from both the Kolmogorov-Smirnov test (KS-test) and a one-way Analysis of Variance (ANOVA) F-test. Additionally, DRA assessment methods are presented using both qualitative (subjective) and quantitative selection. As considered previously for RF-DNA applications [11] GRLVQI feature relevance ranking was used to provide the *post-classification* DRA performance baseline using a full-dimensional feature set. This baseline was used for assessing performance of the *pre-classification* DRA feature selection methods. While feature selection using *p*-values for ranking is quite common [3, 12-16], herein it is illustrated that test statistic values offer many advantages for feature relevance ranking.

The paper is organized as follows. Section II provides a summary of ZigBee RF-DNA Fingerprinting as adopted from [9], and the GRLVQI classifier of [11]. Section III addresses DRA methods, followed by Section IV and Section V which

Research sponsored in part by the Sensors Directorate, Air Force Research Laboratory, Wright-Patterson AFB, Dayton OH . "The views expressed in this article are those of the authors and do not reflect the official policy of the United States Air Force, Department of Defense, or the U.S. Government.

present PHY security contributions, including: 1) introduction of F-test DRA, 2) quantitative vs. qualitative DRA, 3) *p*-values vs. test statistics for feature selection resolution, and 4) DRA performance evaluation driven by device classification and ID verification accuracy. Preliminary Z-Wave device classification is addressed in Section VI and illustrates potential for extended applicability while motivating subsequent research.

II. BACKGROUND

ZigBee devices transmit a structured burst type signal based on a PHY Protocol Data Unit (PPDU) and containing a consistent 40-bit binary zero string Synchronization Header Response (SHR), a defined 8-bit PHY Header Response (PHR), and a variable length 'payload' (PSDU) which consists of a MAC sublayer frame [3, 12, 13]. The RF-DNA process has been previously demonstrated using the entire SHR response to generate RF-DNA fingerprints [3, 12, 13].

A. ZigBee Signal Collection

ZigBee emissions from four Texas Instruments CC2420 2.4GHz transmitters were used here for like-model assessments [3, 12, 13] and device differences are therefore attributable to production variation. The ZigBee devices transmitted at 2.4GHz, within the Agilent collection receiver range of [20.0MHz to 6.0GHz]. As described by [3, 12, 13], burst signals (1000 SHR responses) were collected under three different operating conditions: 1) in a Ramsey STE3000B RF anechoic chamber (CAGE), 2) along a direct line-of-sight (LOS) in an office hallway, and 3) through an office wall (WALL). Additive White Gaussian Noise (AWGN) was combined with collected emissions to achieve $SNR \in [0\ 30]$ dB and simulate varying channel conditions [3, 12, 13].

B. RF-DNA Statistical Fingerprint Generation

RF-DNA Fingerprinting enables device discrimination using differences in transmitted signal characteristics among various devices [9]. RF-DNA fingerprints have been shown to be reliable and accurate for various devices and standards, see [9]. Herein, RF-DNA for intra-device variations of ZigBee devices is considered, extending [3, 12, 13].

Consistent with prior work, [3, 10-13], N_S=3 RF-DNA fingerprints features of variance (σ^2), skewness (γ), and kurtosis (κ) were computed for N_R=80 regions of interest within N_C=3 ZigBee instantaneous time domain amplitude (a), phase (ϕ), and frequency (f) responses. RF-DNA fingerprints were generated by 1) dividing each of the signal responses into N_R contiguous and equal length bins, 2) calculating N_S features for each bin, plus an additional set for the entire response ($N_R + 1$ total bins), and 3) computed features into regional fingerprint vectors as,

$$F_{Ri} = [\sigma_{Ri}^2, \gamma_{Ri}, \kappa_{Ri}\ ]_{1\times 3}, \quad (1)$$

where $i = 1,2,\ldots,N_R + 1$ [3]. A fingerprint vector for each of the N_C characteristics is formed from (1) as,

$$F^C = \left[F_{R_1} \vdots F_{R_2} \cdots F_{R_{(N_R+1)}}\right]_{1\times N_S(N_R+1)}, \quad (2)$$

which are concatenated to form the final fingerprint vector:

$$\boldsymbol{F} = \left[\boldsymbol{F^a} \vdots \boldsymbol{F^\phi} \vdots \boldsymbol{F^f}\right]_{1\times N_S(N_R+1)\times N_C}. \quad (3)$$

For ZigBee device discrimination assessments, a total of N_F=729 features are computed with N_{TRN}=1,500 Training (TNG) and N_{TST}=1,500 Testing (TST) observations; given such a large amount of data DRA is therefore of interest.

C. GRLVQI Classifier Model Development

The GRLVQI classifier method is employed herein, as in [11]. GRLVQI is an extension of Kohonen's self-organizing maps that employs gradient descents, relevance learning, and a sigmoid cost function to train prototype vectors to a given class label [11]. For all devices used herein, prior probabilities were considered equal between devices and the GRLVQI classifier model was created as described in [3, 12].

D. Classification and Verification

The PHY device identification process follows general biometric identification and digital forensic processes, e.g. [17-19]. Consistent with [9], classification performance, "one versus many," is evaluated by considering average percent correct training and testing percent correct (%*C*) versus SNR level for authorized devices [12]. DRA "gain" (dB) over baseline performance at an arbitrary *%C*=90% benchmark for both the TNG and TST sets. Gain is defined here as the reduction in required *SNR*, expressed in dB, for two methods to achieve the same *%C* when compared to the full dimensional baseline.

Device ID verification is a "one versus one" (claimed vs actual) ID assessment where a trained classifier is considered along with probability mass functions (PMFs) for both authorized and rogue devices [12]. Two relevant performance metrics include True Verification Rate (TVR) for authorized devices and Rogue Rejection Rate (RRR) for rogue devices [12]. Binary grant/deny network access decisions are based on verification criteria that includes TVR>90% and RRR>90%.

III. DIMENSIONALITY REDUCTION ANALYSIS (DRA)

DRA consists of both *feature selection*, selecting subsets of existing features, and *feature extraction*, involving data transformation and selection of transformed features [20]. Inherently, the RF-DNA process itself was *feature extraction* and now selecting salient features is of interest. DRA assessment, determining the amount of data to retain, is another important DRA aspect. In all DRA cases, only the TNG feature set was used, thus preserving the TNG/TST feature set sequestration. In all DRA cases, only the TNG feature set was used, thus preserving the TNG/TST feature set sequestration.

A. Feature Selection

The KS-test, F-test, and GRLVQI relevance ranking are the feature selection methods considered herein. The KS-test and F-test are *pre-classification* approaches that consider data distribution aspects, while GRLVQI relevance ranking is a *post-classification* approach based on the contribution of each feature to the full dimensional baseline classifier model. KS-test summed *p*-values were considered and compared for RF-DNA feature selection previously [3, 12, 13]; KS-test statistic values, F-test *p*-values and F-test statistic values were not previously explored for RF-DNA feature ranking application.

1) GRLVQI Relevance

GRLVQI relevance scores, γ, provide a direct indication of feature contribution to classifier development [11, 12]. Higher γ values indicate a feature provides increased class separation, in a GRLVQI classifier, and lower values indicate less class separation [12]. Prior work [12] demonstrated γ-values offering comparable performance to KS-test p-value ranking for ZigBee feature selection with multiple discriminant analysis (MDA).

2) Kolmogorov-Smirnov Test (KS-Test)

The two sample KS-test was employed for ZigBee RF-DNA feature selection in [3, 12, 13]. The KS-test is a distribution-based goodness-of-fit test that considers two sample data vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ and computes a KS-test statistic,

$$\mathrm{KS} = \max\left(|F_1(\mathbf{x}) - F_2(\mathbf{x})|\right), \tag{4}$$

where $F_1(\mathbf{x})$ is the proportion of $\boldsymbol{x}_1$ values less than or equal to $\boldsymbol{x}$ and $F_1(\mathbf{x})$ is the proportion of $\boldsymbol{x}_2$ values less than or equal to $\boldsymbol{x}$ [21]. KS-test p-values are computed against a null distribution, with an implicit null hypothesis that $\boldsymbol{x}_1$and $\boldsymbol{x}_2$ are from the same distribution [12, 13]. For the KS-test, data degrees of freedom (*DoF*) and a null distribution are used to compute p-values, with p-values of 0 possible [21].

For classifier development one should logically seek $\boldsymbol{x}_1$and $\boldsymbol{x}_2$ from different distributions to aid group discrimination [3, 12, 13]. For multiple class problems, such as the ZigBee dataset, pairwise KS-test statistic values are computed for each feature and then combined through summation or averaging. Summed p-values were previously considered for featuring [3, 12, 13]; herein test statistic values are also considered.

3) One-way ANOVA F-Test

General linear models, e.g. ANOVA and linear regression, work to understanding variability of data through sums of squares [22]. The F-test is a heuristic to compute the test statistic for general linear model, with

$$\mathrm{F_{stat}} = \left(\frac{\mathrm{SSF}}{\mathrm{DoF}}\right)/\mathrm{MSE}, \tag{5}$$

where the sum-of-squares for a factor (SSF) and mean squared error (MSE) are from a computed linear ANOVA model involving the data and class labels [22]. Significance for ANOVA problems is determined through F-test p-values computed from a continuous normal distribution with the null that the means of all classes are the same [22]. Feature ranking by F-test statistic values considers the philosophy that higher F_{stat} values indicate a given feature offers higher discrimination between groups [23].

B. Dimensionality Assessment (DA)

Dimensionality assessment involves selecting an appropriate quantity of features. Three approaches to DA are considered: 1) subjective/qualitative, 2) quantitative p-values and 3) quantitative data covariance matrix eigenvalues.

1) Qualitative DRA Assessment

Prior RF-DNA research in [3, 12, 13] examined qualitative DRA methods of subjective "best guesses" for selecting the number of features to retain. For ZigBee, a qualitatively selected N_F = [25, 50, 100, 200, 243] were examined with N_F=50 shown to offer sufficient classification performance [12]. Therefore, N_F=50 is considered herein for comparison.

2) Quantiative: p-value DRA Assessment

One quantitative DRA assessment method involves selecting N_F from p-value significance [14, 15]. Significance levels of [0.1%, 1%, 5%, 10%] are commonly used and justifiable in many cases [24]. Table I presents the indicated number of features to retain for these significance levels using the F-test and KS-test at *SNR*=[0, 10, 18, 30] dB. Comparing Table I with results in [3] indicates that p-value DRA assessment over-estimates the number of features to retain since phase (ϕ) features, N_F=243 herein, are known to offer performance comparable to the baseline. Therefore, p-value dimensionality assessment appears neither appropriate or is considered for ZigBee RF-DNA data.

TABLE I
DIMENSIONALITY ASSESSMENT BY SIGNIFICANCE LEVEL

SNR	METHOD	SIGNIFICANCE LEVEL			
		0.1%	***1%***	***5%***	***10%***
0dB	F-TEST	196	264	350	402
	KS-TEST (ΣP-VALUES)	37	74	130	160
10dB	F-TEST	589	639	674	688
	KS-TEST (ΣP-VALUES)	337	414	512	557
18dB	F-TEST	706	713	720	722
	KS-TEST (ΣP-VALUES)	666	692	711	716
30dB	F-TEST	718	725	727	728
	KS-TEST (ΣP-VALUES)	727	729	729	729

3) Quantiative: Eigenvalue DRA Assessment

Quantitative dimensionality selection based on the data's covariance matrix eigenvalues aim to understand the intrinsic dimensionality of a dataset [22]. Two methods are considered: Kaiser's Criterion (K1) and the Maximum Distance Secant Line (MDSL). The TNG feature set covariance matrix at *SNR*=18 dB and baseline *%C* = 90% was considered.

K1 estimates dimensionality by considering the number of covariance eigenvalues greater than the mean covariance eigenvalue [22]. K1 at *SNR*=18 dB retains N_F=123 features. Cattell's Scree Test involves visually examining a Scree plot (eigenvalues plotted versus rank order) and selecting eigenvalues above the inflection point, the proverbial 'knee in the curve' [22]. MDSL [25] both removes subjectivity and automates Cattell's Scree Test by 1) creating a line between the first and last eigenvalue and 2) finding the point with the largest perpendicular distance from this line, i.e. the inflection point. Using MDSL at *SNR*=18 dB retains N_F=26 features.

IV. FEATURE SELECTION BY TEST STATISTIC OR P-VALUES

Consistency is not seen in p-value or test statistic feature ranking, with both test statistic [23, 26, 27] and p-values [1, 3, 13, 14, 28, 29] used in various applications. Thus, a phenomenological understanding of test statistic and p-values is needed to better understand appropriate uses of both.

Test statistics (magnitudes) are commonly converted to p-values (probabilities) to assess statistical significance related to the probability of a given outcome given the *DoF*, hypothesis test, and a reference distribution [30]. While p-values are related to test statistic values, various issues exist in using p-

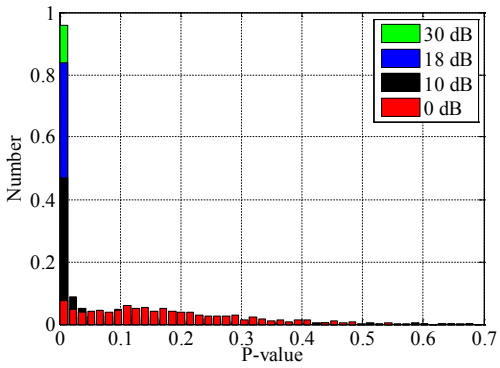

Fig. 1. Normalized histogram of *summed* pairwise KS-test *p*-values for full dimensional (N_F = 729) TNG features for varying *SNR* = [0, 10, 18, 30] dB.

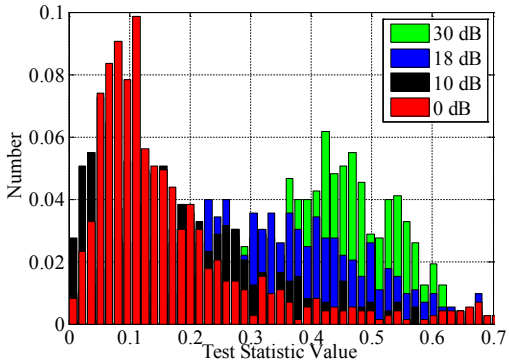

Fig. 2. Normalized histograms of *mean* pairwise KS-test statistic values for full dimensional (N_F = 729) TNG features for *SNR* = [0, 10, 18, 30] dB.

values, including: 1) computing *p*-values involves an implicit distributional assumption whereas test statistics are often only ratios; 2) *p*-values generally involve a nonlinear transformation of test statistic values; 3) *p*-values imply a hypothesis test, but these are not always stated in feature selection, e.g. [12, 13]; 4) *p*-values involve an additional computational step; and 5) *p*-values often converge on 0 as sample size increases [27, 30].

Table II considers test statistics and *p*-values (rank ordered by test statistic) for both F-test and KS-test at SNR=18dB. The variance of test statistic values and *p*-values is also presented. Noticeably, many *p*-values are below the approximate decimal machine precision value, per [31], which indicates that these values are notionally very similar and similarly close to 0. Ranking values equal to or equivalent to 0 logically could be ineffective when selecting a low number of features. Table III further illustrates that *p*-values are seen to trend towards machine precision as *SNR* increases, indicating that increasing signal strength corresponds to increasing significance. Similar issues with *p*-values trending towards 0 were noted in [27, 30].

Normalized histograms (unit area, identical bin centers and widths) for the KS-test summed *p*-values, Fig. 1, and mean test statistic values, Fig. 2, are presented for *SNR* = [0, 10, 18, 30] dB operating points. Similar to Tables I-III, Fig. 1 shows that as noise diminishes, features may appear relevant and approach a *p*-value of 0; therefore *p*-value ranking presents situations where most features are viewed as equally significant. The advantage of using test statistic values is illustrated in Fig. 2; while resolution is lost in Fig. 1, the *mean* test statistic values in Fig. 2 offer a more consistent approach for finding and selecting features. F-test *p*-values and test statistic values show similar distributions (as indicated in Tables I-III.

TABLE II

P-VALUES VS TEST STATISTICS AT SNR = 18dB ORDERED BY DECREASING F-TEST AND KS-TEST TEST STATISTIC VALUES

FEATURE	F-test		KS-TEST	
	Test Statistic	***p-value****	***Summed test statistic***	***summed p-value****
1	542.64	N/A	2.7349	N/A
2	471.78	N/A	2.6487	0
3	432.97	N/A	2.5685	N/A
4	424.26	N/A	2.3065	N/A
5	420.74	N/A	,2.2999	N/A
⋮	⋮	⋮	⋮	⋮
728	0.280	0.839	0.1260	0.9297
729	0.043	0.988	0.1179	1.2285
VARIANCE	6,324.8	0.0094	0.2417	0.0646

*N/A indicates a value at or below an approximate decimal value of general machine precision of 2.22×10^{-16} [31], for the 64-bit PC used herein. Zero variance was assumed for N/A entries.

TABLE III

ZIGBEE *P*-VALUES LESS THAN OR EQUAL TO MACHINE PRECISION

	SNR (dB)			
	0	***10***	***18***	***30***
F-TEST	1.65%	44.99%	78.6%	87.11%
KS-TEST (ΣP-VALUES)	0%	16.74%	54.46%	93.14%

V. ZIGBEE RESULTS

A. Classification Performance

DRA classification performance was compared against full-dimensional baseline (N_F=729) performance using each DRA method considered. Resultant DRA TST classification performance for N_F = [17, 50, 123, 729] DRA feature sets is presented in Fig. 3, with Table IV showing relative DRA "gain" (dB) for both TNG and TST classification.

Most noticeable in Fig. 3 is an overall trend with most DRA methods offering similar curves and with larger N_F values achieving better performance. While the mean test statistic and summed *p*-value ranking methods show comparable performance, the mean KS-test statistic is the only method that achieved positive gain. Additionally, the test statistic approaches consistently outperform the *p*-value approaches, particularly for N_F=17. Of interest in Fig. 3 is that retaining the top N_F=50 or N_F=123 features from F-test, GRLVQI rankings, or KS-test rankings offer comparable performance to the full-dimensional (N_F=729) baseline performance for S*NR*≥8 dB, as noted in [12] for N_F=50.

B. Verification Performance

Full-dimensional baseline ID verification performance was considered for the GRLVQI classifier using the ZigBee dataset for both authorized and unauthorized rogue devices. Baseline performance was determined as TVR=25% for authorized and RRR=66.67% for rogue devices. For DRA classification, the results in Table IV shows that most DRA methods offer poor verification performance. The only exception is the classifier model based GRLVQI relevance rankings, where DRA offers an improvement over the baseline. These results differ from those of [12], which examined an MDA classifier, and could likely be a result of GRLVQI relevance being more applicable to results from the nonlinear GRLVQI classifier.

TABLE IV
ZIGBEE DEVICES: RELATIVE DRA GAIN* (DB) OVER FULL-DIMENSIONAL BASELINE PERFORMANCE AT %C = 90% ACCURACY.

DRA METHOD	KS TEST STATISTIC			KS ΣP-VALUE		
N_F	***17***	***50***	***123***	***17***	***50***	***123***
TNG	-5.00	-0.93	-0.16	-6.39	-1.22	0.00
TST	-6.32	-1.01	**+0.44**	-6.97	-1.33	**-0.70**
DRA METHOD	F TEST STATISTIC			F TEST P-VALUE		
N_F	***17***	***50***	***123***	***17***	***50***	***123***
TNG	-6.83	-2.00	-0.73	-10.92	-1.93	-0.57
TST	-7.59	-1.74	**-0.91**	-10.36	-1.7	**-1.18**
DRA METHOD	GRLVQI					
N_F	***17***	***50***	***123***			
TNG	**-1.44**	-0.60	-0.16			
TST	**-2.61**	**-0.63**	-0.73			

*Bold indicates best or worst results for a given N_F

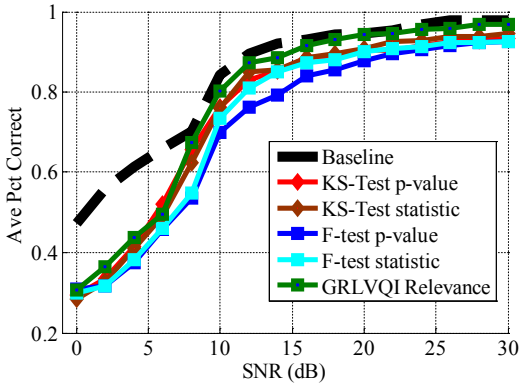

Fig. 3. ZigBee GRLVQI *Testing (TST)* classification performance for full-dimensional (N_F=*729*) and DRA (N_F=*17*) feature sets.

VI. Z-WAVE PRELIMINARY RESULTS

Z-Wave is considered less secure than ZigBee [32]. Three Aeotec Z-Stick S2 transmitters were thus considered as an extension to this research. For analysis, a total of 230 Z-Wave preambles (the first 8.3 ms at 2Msps) collected for N_C=3 devices. Transmission detection from background noise was accomplished through amplitude-based leading edge detection with a -6 dB threshold. The Z-Stick S2 transmitters were located 10 cm from a vertically-oriented LP0410 log periodic antenna, connected via a Gigabit Ethernet cable directly to the USRP-2921 RF input. The resultant collected was SNR=24 dB and like-filtered. AWGN was added to collected signals to achieve desired operating points of SNR∈[0 24] dB in 2 dB steps. A total of 189 RF-DNA features were computed for Z-Wave with N_S=3, N_R=20, N_C = 3, N_{TRN}=115, and N_{TST}=115.

Z-Wave RF-DNA fingerprint features were examined via the F-test and KS-test feature relevance ranking approaches. Table VI presents the number of *p*-values less than or equal to machine precision, illustrating on Z-Wave devices that *p*-value ranking again presents many unusable results.

The GRLVQI classifier achieved *%C*=90% at SNR=20dB, the operating point used for DRA dimensionality assessment. No prior research exists on DRA or RF-DNA for Z-Wave, therefore the quantitative dimensionality approaches in Section III.B.3 were considered with K1 indicating N_F=7 and MDSL indicating N_F=34. Overall results in Table VII and Fig. 4 show that all DRA approaches consistently provide positive dB gain over the full-dimensional baseline.

TABLE V
DRA VERIFICATION PERFORMANCE* FOR AUTHORIZED AND ROGUE DEVICES AT SNR = 18DB FOR TVR≥ 90% AND RRR≥90%.

DRA METHOD	KS TEST STATISTIC			KS ΣP-VALUE		
N_F	***17***	***50***	***123***	***17***	***50***	***123***
TVR	0%	0%	0%	0%	0%	0%
RRR	8.33%	8.33%	0%	**52.8%**	2.78%	0%
DRA METHOD	F TEST STATISTIC			F TEST P-VALUE		
N_F	***17***	***50***	***123***	***17***	***50***	***123***
TVR	0%	0%	0%	**25%**	0%	0%
RRR	8.33%	5.56%	0%	38.9%	19.4%	0%
DRA METHOD	GRLVQI					
N_F	***17***	***50***	***123***			
TVR	**25%**	**50%**	**50%**			
RRR	**52.8%**	**66.7%**	**72.2%**			

*Bold indicates best performance for a given N_F

TABLE VI
PERCENTAGE OF Z-WAVE P-VALUES LESS THAN OR EQUAL TO MACHINE PRECISION AS A FUNCTION OF SNR

	SNR (dB)			
	0	***10***	***20***	***24***
F-TEST	0%	0%	57.14%	60.85%
KS-TEST (ΣP-VALUES)	0%	0%	3.7%	12.17%

TABLE VII
Z-WAVE RELATIVE DRA GAIN (DB) OVER BASELINE PERFORMANCE AT %C = 90% ACCURACY*

DRA METHOD	KS TEST STATISTIC		KS ΣP-VALUE	
N_F	***7***	***34***	***7***	***34***
TNG	+1.92	+1.83	+1.7	+1.64
TST	**+1.79**	**+0.63**	+1.79	+1.37
DRA METHOD	F TEST STATISTIC		F TEST P-VALUE	
N_F	***7***	***34***	***7***	***34***
TNG	**+1.95**	**+2.13**	**+2.22**	+1.82
TST	+0.79	+0.54	+0.63	+2.16
DRA METHOD	GRLVQI			
N_F	***7***	***34***		
TNG	+0.63	+0.24		
TST	+1.41	+0.18		

*Bold indicates best or worst results for a given N_F

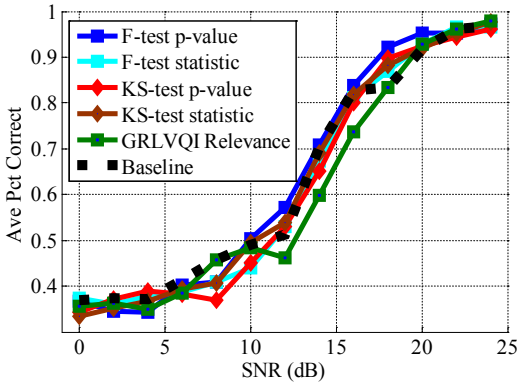

Fig. 4. Z-Wave GRLVQI *Testing (TST)* classification performance for full-dimensional (N_F=*189*) and DRA (N_F=*34*) feature sets.

VII. Summary and Conclusions

The work provided four contributions for improving ZigBee PHY device identification using DRA selected RF-DNA features: 1) the introduction of F-test for RF-DNA DRA; 2) comparative test statistic and *p*-value assessment showing multiple benefits of test statistic values over *p*-values, with test statistic values shown to not converge on any specific number and offer a more natural tool for comparison than *p*-values; 3) introduction of quantitative DRA assessment for RF-DNA features; and 4) a performance comparisons for five DRA approaches using a GRLVQI classifier. Preliminary results using Z-Wave devices showed similar implications.

Both ZigBee and Z-Wave results show that a properly selected feature set provides better device classification and ID verification performance than the full-dimensional baseline feature set. ZigBee results with N_F=123 also achieved better classification performance than earlier N_F=243 results in [12]. DRA was explored with a goal toward selecting 1) robust salient features and 2) an appropriate number of features to maintain classifier integrity. Additionally, the results collectively illustrated that 1) DRA does not always imply classification/ verification performance improvements and 2) DRA closest to the computations used for classifier model development is beneficial for verification performance.

VIII. Bibliography

[1] B. Ramsey, B. Mullins, R. Speers and K. Batterton, "Watching for weakness in wild WPANs," *Military Commun. Conf. (MILCOM),* pp. 1404-1409, 2003.

[2] A. Coustasse, S. Tomblin and C. Slack, "Impact of radio-frequency identification (RFID) technologies on the hospital supply chain: a literature review," *Perspectives in Health Inform. Manag.,* vol. 10, 2013.

[3] B. Ramsey, M. Temple and B. Mullins, "PHY foundation for multi-factor ZigBee node authentication," *Global Commun. Conf. (GLOBECOM),* pp. 795-800, 2012.

[4] R. Begley, "Development of Autonomous Railcar Tracking Technology Using Railroad Industry Radio Frequencies," *Research Opportunities in Radio Frequency Identification Transportation Applicat.,* p. 59, 2006.

[5] Y. Zatout, "Using wireless technologies for healthcare monitoring at home: A survey," *Int. Conf. e-Health Networking, Applicat. and Services (Healthcom),* pp. 383-386, 2012.

[6] J. Wright, "KillerBee: practical zigbee exploitation framework," in *11th ToorCon Conf.*, San Diego, 2009.

[7] Y. Sheng, K. Tan, G. Chen, D. Kotz and A. Campbell, "Detecting 802.11 MAC layer spoofing using received signal strength," *27th Conf. Comput. Commun.,* 2008.

[8] B. Danev et al., "On physical-layer identification of wireless devices," *ACM Computing Surveys,* vol. 45, no. 1, 2012.

[9] W. Cobb et al., "Physical layer identification of embedded devices using RF-DNA fingerprinting," *Military Commun. Conf. (MILCOM),* pp. 2168-2173, 2010.

[10] M. Williams, M. Temple and D. Reising, "Augmenting bit-level network security using physical layer RF-DNA fingerprinting," *Global Commun. Conf. (GLOBECOM),* pp. 1-6, 2010.

[11] P. Harmer, D. Reising and M. Temple, "Classifier selection for physical layer security augmentation in Cognitive Radio networks," *Int. Conf. Commun. (ICC),* pp. 2846-2851, 2013.

[12] C. Dubendorfer, B. Ramsey and M. Temple, "ZigBee device verification for securing industrial control and building automation systems," *Int. Conf. Critical Infrastructure Protection (IFIP13),* vol. 417, pp. 47-62, 2013.

[13] C. Dubendorfer, B. Ramsey and M. Temple, "An RF-DNA verification process for ZigBee networks," *Military Commun. Conf. (MILCOM),* pp. 1-6, 2012.

[14] T. Wu, J. Duchateau, J.-P. Martens and D. van Compernolle, "Feature subset selection for improved native accent identification," *Speech Commun.,* vol. 52, no. 2, pp. 83-98, 2010.

[15] A. Haury, P. Gestraud and J.-P. Vert, "The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures," *PLoS ONE,* vol. 6, no. 12, 2011.

[16] T. Kind, V. Tolstikov, O. Fiehn and R. Weiss, "A comprehensive urinary metabolomic approach for identifying kidney cancer," *Analytical Biochemistry,* vol. 363, pp. 185-195, 2007.

[17] S. Prabhakar, S. Pankanti and A. K. Jain, "Biometric recognition: Security and privacy concerns," *IEEE Security and Privacy,* pp. 33-42, March/April 2003.

[18] A. King and P. Wahjudi, "Dynamic Free Text Keystroke Biometrics System for Simultaneous Authentication and Adaptation to User's Typing Pattern," *J. Manage. & Eng. Integration,* vol. 6, no. 2, pp. 86-93, 2013.

[19] D. W. Baker et al., "Digital evolution: History, challenges and future directions for the digital and multimedia sciences section," *Forensic Sci.: Current Issues, Future Directions,* pp. 252-291, 2013.

[20] A. K. Jain, R. P. Duin and J. Mao, "Statistical Pattern Recognition: a Review," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. vol. 22, no. no. 1, pp. 4-37, Jan. 2000.

[21] W. J. Conover, *Practical Nonparametric Statistics*, 2nd ed., New York: John Wiley & Sons, 1980, pp. 344-385.

[22] W. R. Dillon and M. Goldstein, *Multivariate Analysis Methods and Applications*, New York: John Wiley & Sons, 1984.

[23] J. D. Habbema and J. Hermans, "Selection of variables in discriminant analysis by F-statistic and error rate," *Technometrics,* vol. 19, no. 4, pp. 487-493, 1977.

[24] M. Cowles and C. Davis, "On the Origins of the .05 Level of Statistical Significance," *Amer. Psychologist,* vol. 37, no. 5, pp. 553-558, 1982.

[25] R. Johnson, J. Williams and K. Bauer, "AutoGAD: An improved ICA-based hyperspectral anomaly detection algorithm," *IEEE Trans. Geosci. Remote Sens.,* vol. 51, no. 6, pp. 3492-3503, 2013.

[26] C. J. Huberty and J. M. Wisenbaker, "Variable importance in multivariate group comparisons," *J. Educ. Stat.,* vol. 17, no. 1, pp. 75-91, 1992.

[27] A. Cord, C. Ambroise and J.-P. Cocquerez, "Feature selection in robust clustering based on Laplace mixture," *Pattern Recognition Letters,* vol. 27, no. 6, pp. 627-635, 2006.

[28] P. Radivojac, Z. Obradovic, A. K. Dunker and S. Vucetic, "Feature selection filters based on the permutation test," *Machine Learning: ECML 2004,* pp. 334-346, 2004.

[29] K. Schmidt, T. Behrens and T. Scholten, "Instance selection and classification tree analysis for large spatial datasets in digital soil mapping," *Geoderma,* vol. 146, no. 1-2, pp. 138-146, 2008.

[30] L. G. Halsey et al., "The fickle P value generates irreproducible results," *Nature Methods,* vol. 12, no. 3, pp. 179-185, 2015.

[31] C. Moler, "Floating Points," *MATLAB News and Notes,* pp. 1-3, 1996.

[32] M. Knight, "How safe is Z-Wave?[Wireless standards]," *Computing and Control Eng.,* vol. 17, no. 6, pp. 18-23, 2006.