data

MDPI

Data Descriptor

A Database for the Radio Frequency Fingerprinting of Bluetooth Devices

Emre Uzundurukan [1], Yaser Dalveren [1,2,*] and Ali Kara [3]

[1] Department of Avionics, Atilim University, 06830 Ankara, Turkey; emre.uzundurukan@atilim.edu.tr
[2] Department of Electronic Systems, Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology, 2815 Gjøvik, Norway
[3] Department of Electrical Electronics Engineering, Atilim University, 06830 Ankara, Turkey; ali.kara@atilim.edu.tr
* Correspondence: yaser.dalveren@ntnu.no

Received: 19 May 2020; Accepted: 19 June 2020; Published: 21 June 2020

Abstract: Radio frequency fingerprinting (RFF) is a promising physical layer protection technique which can be used to defend wireless networks from malicious attacks. It is based on the use of the distinctive features of the physical waveforms (signals) transmitted from wireless devices in order to classify authorized users. The most important requirement to develop an RFF method is the existence of a precise, robust, and extensive database of the emitted signals. In this context, this paper introduces a database consisting of Bluetooth (BT) signals collected at different sampling rates from 27 different smartphones (six manufacturers with several models for each). Firstly, the data acquisition system to create the database is described in detail. Then, the two well-known methods based on transient BT signals are experimentally tested by using the provided data to check their solidity. The results show that the created database may be useful for many researchers working on the development of the RFF of BT devices.

Dataset: http://doi.org/10.5281/zenodo.3876140

Keywords: Bluetooth; classification; data acquisition; emitter identification; radio frequency fingerprinting; RF front end

1. Introduction

The development of wireless communication technologies has introduced end users to various external security attacks. For this reason, the protection of wireless communication networks has become an important concern in order to provide data security and user privacy. Currently, one of the ways of achieving this is to employ traditional upper layer security techniques. However, physical layer protection techniques can also be used to protect wireless networks from malicious attacks. Such techniques are based on the use of the fully physical identification of wireless communication devices. This process is also known as radio frequency fingerprinting (RFF). In RFF, the distinctive or unique features of the physical waveforms (signals) emitted by wireless devices are utilized to classify the authorized users. This paves the way for identifying possible threats to the network [1,2].

A typical RFF method is composed of three main stages, namely data acquisition, signal processing and classification. In the literature, high-end receivers [3–5] and low-end receivers [5,6] have been preferred for data acquisition. High-end receivers use higher sampling rates. This, evidently, increases the data size. Hence, an extended memory is highly important to record signals. It should be noted that the sampling rate is one of the most critical parameters that greatly affects the accuracy. Specifically, a higher sampling rate can result in undesired frequency components in the signals, while lower sampling rates can cause the loss of the unique features needed for RFF. To overcome this trade off, it is necessary to use sub-Nyquist rates [7], or downconverters [3]. In this context, an RF front end system can be used for data acquisition [8,9]. Another critical parameter in data acquisition is the device diversity and also the number of signals to be recorded. For a reliable RFF method, the number of devices and captured signals should be kept as high as possible. In the signal processing stage, on the other hand, either the transient or steady-state regions of the transmitted signals are used to extract the distinctive or unique features (so called "RF fingerprints"). The extracted features are then used in the subsequent stage in order to classify the transmitting devices according to their model and manufacturer.

Among the implementation stages of the RFF method, the data acquisition stage plays a critical role as it directly affects the upper bounds of the performance of RFF. This is because even a small error or deficiency, like an insufficient number of devices in the data acquisition, might adversely affect the subsequent stages. Consequently, this leads to a poor device identification ability. Then, the data acquisition stage should be planned carefully in order to provide an accurate, robust and adequate size of database. The purpose of this study is to introduce a database consisting of datasets, including Bluetooth (BT) signals, collected from various smartphones of different brands and models, that were recorded at different sampling frequencies. To this end, the details of the data acquisition system are presented in detail. In addition, the results of two transient-based RFF methods that use the dataset are provided. To the best of our knowledge, this is the first freely available database that enables the testing or developing of RFF methods with various BT devices. Therefore, it is believed that the database would help many researchers from the community of the RFF identification of BT devices.

2. Data Acquisition and Processing

2.1. Data Acquisition System

The dataset was constructed by capturing the emitted signals from each device within a period of several months. As reported in the literature [6–8], the distinctive or unique features of emitted signals (so, the devices) do not vary within short time scales (days, weeks and months). However, for long periods (typically years), it is necessary to study the changes in the emitted signals due to hardware-oriented defects in the devices, which might change the emitted signal's features. On the other hand, there are two more parameters that might affect the unique features of the emitted signals; signal-to-noise ratio (SNR) and interferers in the environment. In this work, these two parameters were considered at all signal collection stages by keeping the distance between the emitting phone and the receiver fixed and also by removing possible interferers in the environment.

In order to capture BT signals, two different systems were used, as described in the following section.

2.1.1. Direct Sampling

The first system for data acquisition is shown in Figure 1. BT signals were directly captured by means of a high sampling rate oscilloscope (Tektronix TDS7404), along with a low-resolution (8 bits) analog-to-digital converter (ADC). As BT devices operate at a ISM2400 band, according to the Nyquist Theorem, a 4.8 Gsps sampling rate is required to capture BT signals. Hence, three alternative sampling rates of the receiver of the oscilloscope were used: 5 Gsps, 10 Gsps and 20 Gsps. A commercial off-the-shelf (COTS) antenna operating at 2.4 GHz was then connected to the oscilloscope. While capturing data, the distance between the antenna and smartphone was fixed at approximately 30 cm. BT signals were captured through the edge detection mode of the oscilloscope. In this way, BT signals with an approximate duration of 10 µs were captured adequately. The data were recorded in text format (.txt). As the BT band covers about 82 MHz, the bandwidth was tuned to 15 MHz–100 MHz.

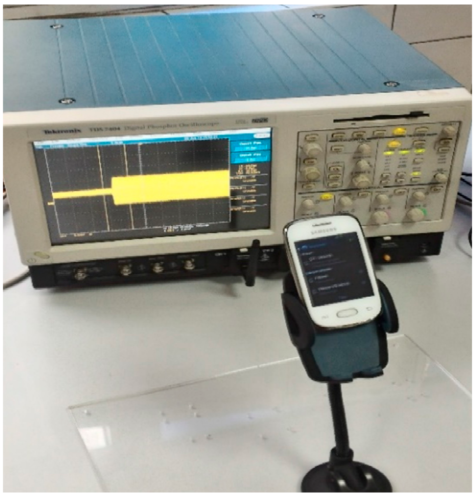


Figure 1. Data acquisition system with direct sampling.

2.1.2. Sampling with RF Front End

The second system for data acquisition is shown in Figure 2. In this system, a modular RF front end system, as in [9], was used. Here, the RF front end was connected to the oscilloscope to record the received signals at lower sampling rates (250 Msps). As mentioned before, this provides several benefits. With the help of this system, the data size along with the computational cost are significantly reduced. Besides this, it also enables researchers to reduce the memory requirements of the data storage equipment. More detailed descriptions and specifications of the RF front end can be found in [9].

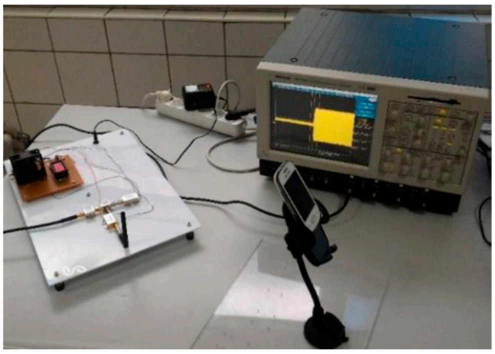


Figure 2. Data acquisition system with radio frequency (RF) front end.

2.2. BT Signal Capturing

BT signals were captured in an isolated laboratory environment, on the second underground floor of a nine-story building in Atilim University, Ankara, Turkey. During the data capturing process, the electronic devices around the system were switched off (possible interferers). To collect BT signals, a total of 27 different smartphones (6 manufacturers with various models) were used. Additionally, two different smartphone series for each model were acquired.

For the first data acquisition system, three datasets were created for the three different sampling rates. The datasets containing BT signals sampled at 5, 10 and 20 Gsps are presented in Table 1. For the second data acquisition system, a dataset containing the BT signals sampled at a rate of 250 Msps is also presented in Table 1. At each dataset, 150 BT signals from each device were recorded. Therefore, a database containing approximately 12,900 recordings from 86 smartphones was created.

Table 1. Datasets, sampling rates and Bluetooth (BT) device sets.

Dataset A (5 Gsps)		**Dataset B (10 Gsps)**		**Dataset C (20 Gsps)**		**Dataset D (250 Msps)**	
Brand	**Model**	**Brand**	**Model**	**Brand**	**Model**	**Brand**	**Model**
Apple	iPhone 5	Apple	iPhone 4s	Apple	iPhone 5s	Apple	iPhone 4s
Apple	iPhone 5s	Apple	iPhone 7	Apple	iPhone 6s	Apple	iPhone 5
Apple	iPhone 6	Apple	iPhone 7 Plus	Apple	iPhone 6s Plus	Apple	iPhone 5s
Apple	iPhone 6s	LG	V20	Apple	iPhone 7	Apple	iPhone 6
LG	G4	Samsung	J7	Huawei	Gr5	Apple	iPhone 6s
Samsung	Note 3	Samsung	Note 2	LG	G4	Apple	iPhone 7
Samsung	S5	Samsung	S7 Edge	Samsung	Note 3	Apple	iPhone 7 Plus
Sony	Xperia M5	Xiaomi	Mi 6	Samsung	S3	LG	G4
				Samsung	S3 Duos	LG	V20
				Samsung	S4	Samsung	J7
				Sony	C4	Samsung	Note 2
						Samsung	Note 3
						Samsung	S5
						Samsung	S7 Edge
						Sony	Xperia M5
						Xiaomi	Mi6

2.3. *Preprocessing*

After recording the signals, it was found that the oscilloscope used in the first data acquisition system generated some undesired signals (spur signals). As shown in Figure 3, these undesired signals were detected at frequencies of 2.5 GHz or above in the datasets of A, B and C. Therefore, a band pass filter (BPF) was used to remove the spur signals. Essentially, this was a standard digital filter allowing only ISM2400 band components.

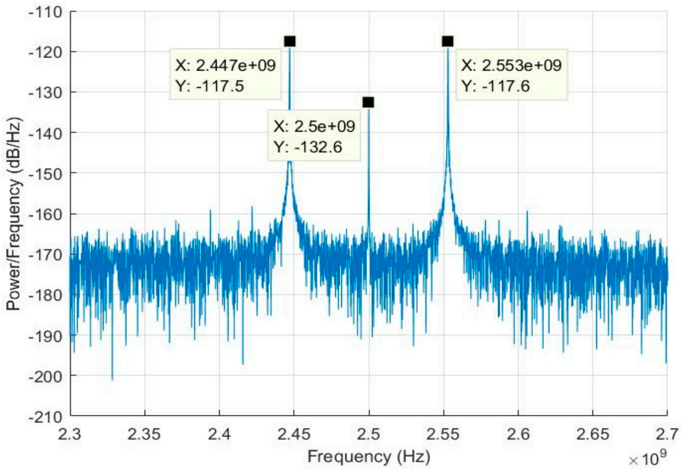
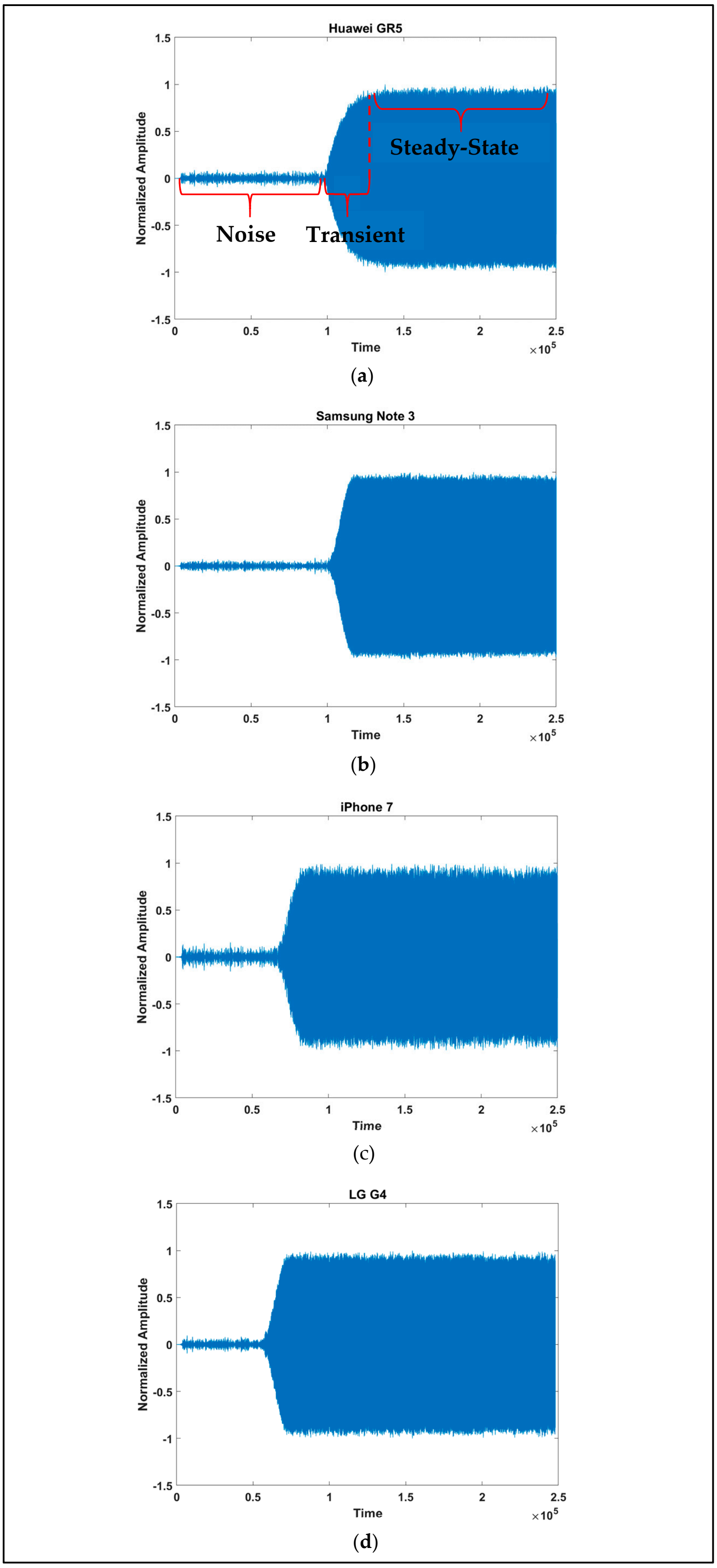


Figure 3. Undesired frequency components (spur signals).

Furthermore, all the signals were normalized for scaling purposes. In this way, the extracted features can represent the generalized features of the emitting devices. As an example, Figure 4 shows the recordings of the BT signals after preprocessing.



(**a**)

(**b**)

(c)

(d)

Figure 4. The recordings of BT signals: (**a**) Huawei GR5; (**b**) Samsung Note 3; (**c**) iPhone 7; (**d**) LG G4.

3. Data Usage Example

The recorded data are a real-valued time series (voltage/time). Firstly, they should be transformed into analytical signals by using the Hilbert transform (HT) [9–12]. Then, a digital downconverter can be used if further down-conversion is needed. Next, the I/Q data can be generated through MATLAB or AWR.

As discussed previously, to classify the authorized users in a network, the distinctive or unique features of the physical waveforms transmitted from a wireless device are utilized in RFF. The distinctive features can be extracted from transient or steady-state regions of the transmitted signals, as shown in Figure 4. Recent studies have shown that most of the distinctive features can be extracted from transient regions [9–12]. To do this, firstly, it is necessary to detect transient signals. Then, from the detected transient signal, the features can be extracted. Mostly, instantaneous signal characteristics [13] and the time–frequency–energy distribution (TFED) [14] are utilized to extract the features. Next, the extracted features are used in the classification of the transmitting devices by brand, model or series. The choice of classifier type might highly affect the RFF's performance. Deep learning (DL) [4], support vector machines (SVM) [14], k-nearest neighbor (KNN) [15] and multiple discriminant analysis (MDA) [16] are well-known for the identification of transmitting devices in RFF.

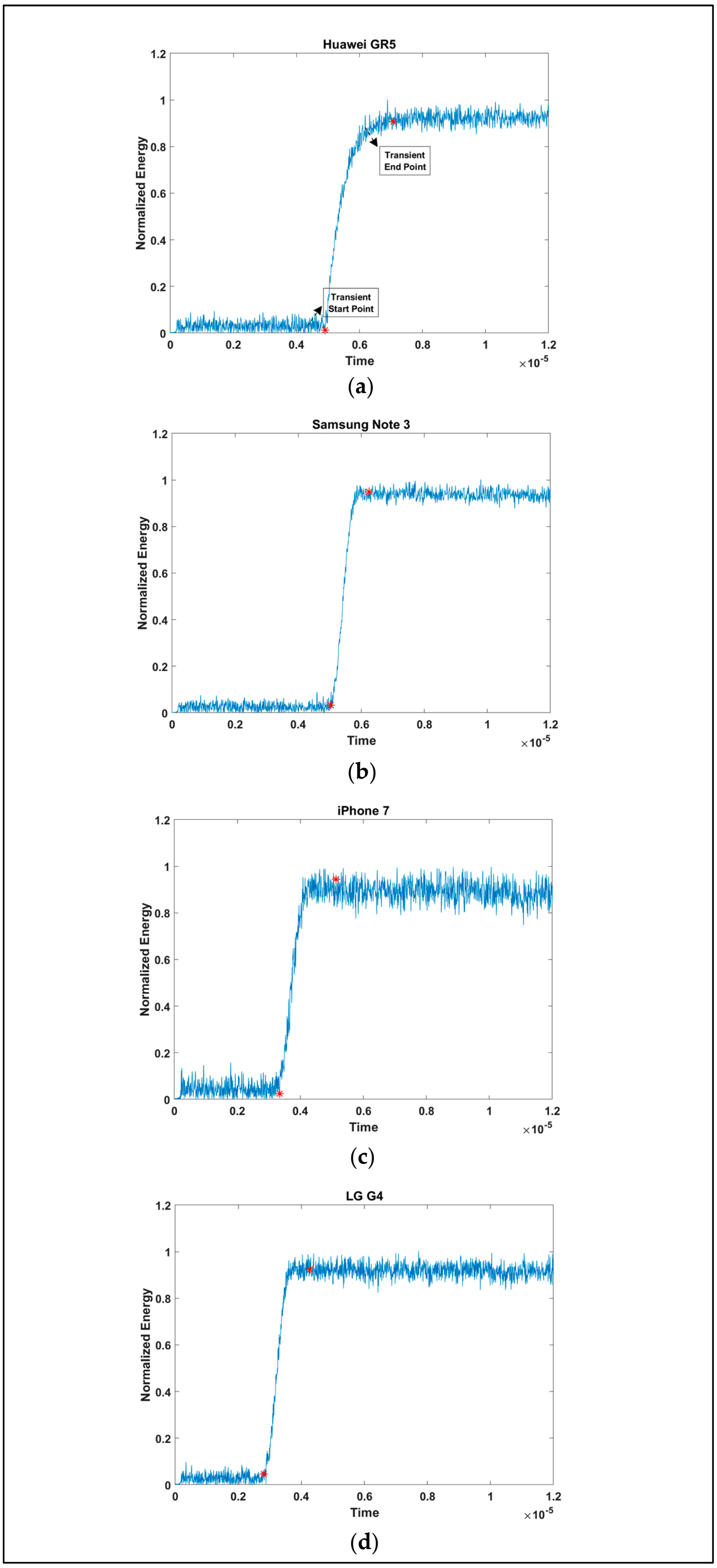
For a demonstration of the use of the dataset, the implementation of RFF is presented here. To this end, two different transient signal-based RFF methods on the basis of instantaneous signal characteristics and TFED features were experimentally tested. These methods are described in the following subsections. Before testing these methods, a transient detection method was employed to detect the transients of the recorded signals [17]. In this method, the energy envelope of the emitted signals was utilized. The transient signals detected from four BT devices are shown in Figure 5. Moreover, the normalized energy of the signals from the same devices are also shown in Figure 6 for comparison.

3.1. Device Identification Using Instantaneous Signal Characteristics

In the transient signal-based RFF method, on the basis of instantaneous signal characteristics, three higher order statistical (HOS) features (skewness, kurtosis and variance) are derived from the signal's characteristics, namely instantaneous amplitude, instantaneous frequency and instantaneous phase. The process of extracting HOS features has been presented in detail in [9–13].

After extracting the HOS features, the next step is the classification of the BT devices (smartphones). The classification performances of two classifiers, the support vector machine (SVM) and neural networks (NN), were examined. The feature set was divided into training and test sets for each BT device in the dataset. Each training set consisted of 120 transient signals (out of 150) while the test set consisted of the remaining 30 transients. The training and test sets were chosen randomly from the dataset.

In the training stage, a non-linear SVM classifier was used as the data were linearly inseparable. To build a non-linear SVM classifier, it is already known that the kernel function enables researchers to map the dataset onto a higher-dimensional vector space. Although there are several types of kernel functions (radial basis, sigmoid and polynomial), the polynomial kernel function (quadratic) was chosen as it provides higher classification accuracy. On the other hand, a multi-layer NN structure for the training of the NN classifier can be found in [9]. As the input layer of the NN structure, ten neurons corresponding to the number of generated signal features were selected. The output layer consisted of sixteen neurons as there were sixteen classes in each dataset. Moreover, two hidden layers with four neurons were used initially. Then, the number of neurons and hidden layers were increased gradually to achieve the best training performance. It was found that three hidden layers with sixty-four neurons for each were sufficient to achieve the targeted performance. Furthermore, the *tansig* function was chosen for the activation of all neurons. For the network training, the number of epoch limits was chosen as 2500. The classification accuracy of the classifiers is given in Table 2.



(**a**)

(**b**)

(**c**)

(**d**)

Figure 5. The detected transient signals: (**a**) Huawei GR5; (**b**) Samsung Note 3; (**c**) iPhone 7; (**d**) LG G4.

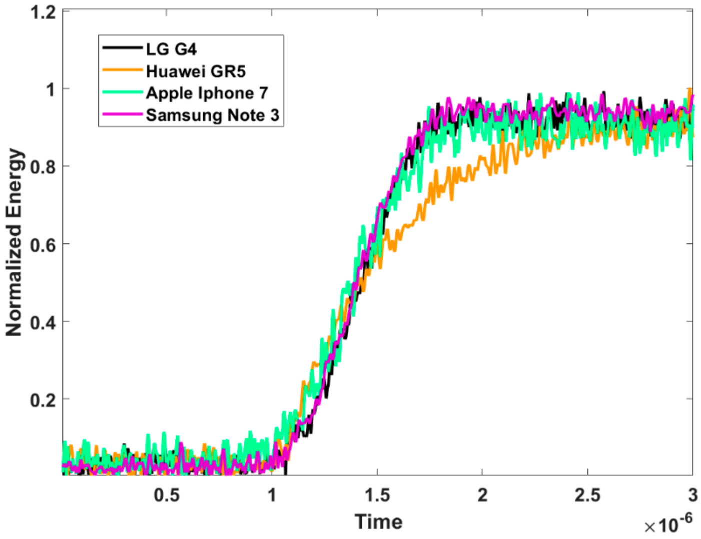


Figure 6. Comparison of the normalized energies of the data collected from Huawei GR5, Samsung Note 3, iPhone 7 and LG G4.

Table 2. Classification performances of support vector machine (SVM) and neural networks (NN).

Dataset	Classifier	Training Accuracy	Test Accuracy
A	SVM	100%	99.2%
B		99.7%	97.9%
C		99.8%	99.0%
D		99.6%	96.9%
A	NN	100%	99.6%
B		100%	97.3%
C		100%	99.4%
D		100%	96.5%

From the results listed in Table 2, the NN classifier initially seems to be overfitted due to the training accuracy. While training the data, initial weights were used randomly at every turn. For this reason, such higher training accuracies would have been expected. If similar data were trained, it would not be possible to achieve such accuracy rates. Here, the network that achieved the highest training accuracy among the trained networks was used. Furthermore, the training and test sets consisted of different data. If there was an overfitting, the overfitted network would memorize only the given training data, and the accuracy of the test data would be much lower. In this context, the high accuracies achieved with both the training and test data indicate that the network was not overfitted.

3.2. Device Identification Using Time Frequency Energy Distribution (TFED) Features

An RFF method based on TFED features is presented in [14]. In this method, features can be extracted by using the Hilbert Huang Transformation (HHT) of transient signals. HHT is simply defined as the calculation of the energy of each frequency component in a certain time resolution. In this context, the device's signal characteristics in terms of both time and frequency can be analyzed easily.

To extract the TFED features, the dataset consisting of BT signals sampled with 20 Gsps was used (dataset C). The features extracted from the TFED of BT signals are listed in Table 3. A feature set created from the features was smoothed by a median filter in order to evaluate the effects of filtering the features on the classification performance. Then, the smoothed and unsmoothed features were employed separately for comparison. Three classifiers, the linear support vector machine (L-SVM), linear discriminant analysis (LDA) and the complex decision tree (CDT) were used for the same feature set. As in the previous section, the feature set was divided into training and testing sets. In the training set, 60 records (out of 150 records) were used, while the remaining 90 records were used in the testing set.

Table 3. Time frequency energy distribution (TFED) feature set.

Feature Number	Feature Name
1	Duration of transient signal
2	Total energy of transient energy
3	Total energy of transient energy envelope
4	Variance of transient energy envelope
5	Standard deviation of instantaneous phase of transient signal
6	Entropy of instantaneous phase of transient signal
7	Length of transient energy distribution
8	Slope of transient energy distribution
9	Variance of summation of transient energy distribution
10	Maximum of summation of transient energy distribution
11	Third order polynomial fitting coefficient of summation of transient energy distribution
12	Maximum of summation of transient energy distribution
13	Variance of summation of transient energy distribution

The CDT classifier consisted of three main nodes: root node, split nodes and leaves or terminal nodes. The root node contains whole data while the split nodes generate internal nodes. CDT is known as a supervised learning algorithm and is widely used in classification processes. To generate the model, it requires high quality training data. The CDT classifier is based on the idea that the data are split into subsets, each of which belongs to a unique class.

On the other hand, the LDA is based on modeling the differences between the data classes. This classifier projects high-dimensional feature vectors onto low-dimensional ones by means of a linear transformation. In this way, the generated vectors provide an efficient separation of the classes.

Finally, linear support vector machines (L-SVM) is a supervised machine learning algorithm which maps input data (x) onto high-dimensional data. For this, it utilizes a mapping function $\varphi(\cdot)$, along with a linear transformation $f(x) = \omega\varphi(x) + b$, where ω and b are the optimized coefficients. With the help of $f(x)$, the data can be separated, from which a hyperplane is generated. Maximizing the margin between the separating hyperplanes results in minimizing the upper bound error, and thus, the structure of the L-SVM is constructed. The L-SVM classifier was employed for the given BT data, and the details are reported in [10,12].

The performances of the classifiers for both the smoothed and unsmoothed features are presented in Table 4.

Table 4. Classification performances of linear support vector machines (L-SVM), linear discriminant analysis (LDA) and complex decision tree (CDT).

Classifier	Smoothed Features Accuracy	Un-smoothed Features Accuracy
L-SVM	99.8%	97.2%
CDT	99.6%	97.6%
LDA	99.7%	91.3%

3.3. Discussion

In order to evaluate the usability of the database, two different transient signal-based RFF methods have been experimentally tested. During the tests, the distinctive features extracted from the transient signals were used in the classification stage to identify the BT devices. The classification performance results of the classifiers prove the robustness of the datasets. Obviously, without an accurate dataset, it is impossible to achieve such classification performance. The results also verify the effectiveness of the data acquisition system. Therefore, the database provided in this study could be valuable for the research community. The database may give greater flexibility to the research community for developing better RFF methods. On the other hand, novel or existing RFF methods can also be implemented with the database. It may also be used in developing and testing transient or steady-state signal detection techniques.

4. Conclusions

This paper is intended to describe a BT signal database which is freely available to the research community for developing RFF methods. The BT database was recorded at an isolated laboratory at Atilim University, Ankara, Turkey. A set of 27 smartphones from various models produced by six manufacturers were used in the data collection. This is a work of a team who dedicated substantial time and effort to generate such an extensive and reliable BT signal database. Even a small mistake in the data acquisition process can adversely affect the following stages in the RFF methods. For this reason, this paper presents not only the database but also the data acquisition methodology for a reliable database of BT signals. Moreover, two well-known RFF methods have been reviewed, and the demonstration results are presented to show the usability of the database. The results of the RRF methods prove the effectiveness of both the acquisition system and the database for the further researching of RFF methods. As a future work, the authors intend to create a new version of the database that might include Wi-Fi signals.

Author Contributions: E.U. contributed to the collection of the data and pre-processing tasks. Y.D. guided the data collection, formulated the structure of the paper and create the draft of the manuscript. A.K. identified problem, coordinated the data collection stages, and evaluated the findings. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bolle, R.M.; Connell, J.H.; Pankanti, S.; Ratha, N.K.; Senior, A.W. *Guide to Biometrics*; Springer Science & Business Media: Berlin, Germany, 2013.
2. Danev, B.; Luecken, H.; Capkun, S.; El Defrawy, K. Attacks on physical-layer identification. In Proceedings of the 3rd ACM Conference on Wireless Network Security, Hoboken, NJ, USA, 22–24 March 2010; pp. 89–98.
3. Talbot, C.M.; Temple, M.A.; Carbino, T.J.; Betances, J.A. Detecting rogue attacks on commercial wireless Insteon home automation systems. *Comput. Secur.* **2017**, *74*, 296–307. [CrossRef]
4. Merchant, K.; Revay, S.; Stantchev, G.; Nousain, B. Deep learning for RF device fingerprinting in cognitive communication networks. *IEEE J. Sel. Top. Signal Process.* **2018**, *12*, 160–167. [CrossRef]
5. Ramsey, B.W.; Stubbs, T.D.; Mullins, B.E.; Temple, M.A.; Buckner, M.A. Wireless infrastructure protection using low-cost radio frequency fingerprinting receivers. *Int. J. Crit. Infrastruct. Prot.* **2015**, *8*, 27–39. [CrossRef]
6. Rehman, S.U.; Sowerby, K.W.; Coghill, C. Radio-frequency fingerprinting for mitigating primary user emulation attack in low-end cognitive radios. *IET Commun.* **2014**, *8*, 1274–1284. [CrossRef]
7. Reising, D.R.; Temple, M.A.; Jackson, J.A. Authorized and rogue device discrimination using dimensionally reduced RF-DNA fingerprints. *IEEE Trans. Inf. Forensics Secur.* **2015**, *10*, 1180–1192. [CrossRef]
8. Rehman, S.U.; Sowerby, K.; Coghill, C. Analysis of receiver front end on the performance of rf fingerprinting. In Proceedings of the 2012 IEEE 23rd International Symposium on Personal, Indoor and Mobile Radio Communications-(PIMRC), Sydney, NSW, Australia, 9–12 September 2012; pp. 2494–2499.
9. Uzundurukan, E.; Ali, A.M.; Dalveren, Y.; Kara, A. Performance analysis of modular RF front end for RF fingerprinting of bluetooth devices. *Wirel. Pers. Commun.* **2020**, *112*, 2519–2531. [CrossRef]
10. Ali, A.M.; Uzundurukan, E.; Kara, A. Assessment of features and classifiers for Bluetooth RF fingerprinting. *IEEE Access* **2019**, *7*, 50524–50535. [CrossRef]
11. Aghnaiya, A.; Ali, A.M.; Kara, A. Variational mode decomposition-based radio frequency fingerprinting of bluetooth devices. *IEEE Access* **2019**, *7*, 144054–144058. [CrossRef]
12. Aghnaiya, A.; Dalveren, Y.; Kara, A. On the performance of variational mode decomposition-based radio frequency fingerprinting of bluetooth devices. *Sensors* **2020**, *20*, 1704. [CrossRef] [PubMed]

13. Klein, R.W.; Temple, M.A.; Mendenhall, M.J. Application of wavelet-based RF fingerprinting to enhance wireless network security. *J. Commun. Netw.* **2009**, *11*, 544–555. [CrossRef]
14. Yuan, Y.; Huang, Z.; Wu, H.; Wang, X. Specific emitter identification based on Hilbert–Huang transform-based time–frequency–energy distribution features. *IET Commun.* **2014**, *8*, 2404–2412. [CrossRef]
15. Rehman, S.U.; Sowerby, K.; Coghill, C. RF fingerprint extraction from the energy envelope of an instantaneous transient signal. In Proceedings of the 2012 Australian Communications Theory Workshop (AusCTW), Wellington, New Zealand, 30 January–2 February 2012; pp. 90–95.
16. Klein, R.W.; Temple, M.A.; Mendenhall, M.J. Application of wavelet denoising to improve OFDM-based signal detection and classification. *Secur. Commun. Netw.* **2010**, *3*, 71–82. [CrossRef]
17. Ali, A.M.; Uzundurukan, E.; Kara, A. Improvements on transient signal detection for RF fingerprinting. In Proceedings of the 2017 25th Signal Processing and Communications Applications Conference (SIU), Antalya, Turkey, 15–18 May 2017; pp. 1–4.