data

MDPI

Data Descriptor

Identifying GNSS Signals Based on Their Radio Frequency (RF) Features—A Dataset with GNSS Raw Signals Based on Roof Antennas and Spectracom Generator

Ruben Morales Ferre *, Wenbo Wang, Alejandro Sanz Abia and Elena Simona Lohan

ITC Faculty, Department of Electrical Engineering, Tampere University, 33720 Tampere, Finland; wenbo.wang@tuni.fi (W.W.); alejandro.sanzabia@tuni.fi (A.S.A.); elena-simona.lohan@tuni.fi (E.S.L.)

* Correspondence: ruben.moralesferre@tuni.fi

Received: 6 January 2020; Accepted: 13 February 2020; Published: 17 February 2020

Abstract: This is a data descriptor paper for a set of raw GNSS signals collected via roof antennas and Spectracom simulator for general-purpose uses. We give one example of possible data use in the context of Radio Frequency Fingerprinting (RFF) studies for signal-type identification based on front-end hardware characteristics at transmitter or receiver side. Examples are given in this paper of achievable classification accuracy of six of the collected signal classes. The RFF is one of the state-of-the-art, promising methods to identify GNSS transmitters and receivers, and can find future applicability in anti-spoofing and anti-jamming solutions for example. The uses of the provided raw data are not limited to RFF studies, but can extend to uses such as testing GNSS acquisition and tracking, antenna array experiments, and so forth.

Dataset: 10.5281/zenodo.3588393

Dataset License: CC BY 4.0

Keywords: Global Navigation Satellite Systems (GNSS); Radio Frequency Fingerprinting (RF FP); spectracom; roof antenna; Galileo; Global Positioning Systems (GPS); machine learning

1. Introduction and Motivation

Over the last decades, Global Navigation Satellite Systems (GNSS) receiver technologies have significantly evolved. Nowadays, many GNSS Software Defined Radio (SDR) solutions are available, where baseband digital representations of raw GNSS data may be used to develop, test, and fine-tune new algorithms. New algorithms relying on raw I/Q GNSS data can serve different purposes, such as enhanced acquisition and tracking solutions or increased resilience (e.g., robustness against interference, multipath, atmosphere effects, interferers such as spoofers and jammers, etc.). The SDR raw data (i.e., I/Q samples) typically require large amounts of memory sizes in order to store it, and only few datasets are available openly in the current literature. To the best of the authors' knowledge, no sizeable dataset is currently available.

It is the main purpose of the authors to provide a set of GNSS SDR raw data with different scenarios in open-access for further testing purposes. The use cases of such data spread in multiple directions, which are left to the choice of the research community. The use case that we give as an example in this paper is a sub-set of Radio Frequency Fingerprinting (RFF) problem, namely a signal classification problem, based on transmitter-specific and receiver-specific features. The transmitter-specific features considered here are the number of GNSS signals and their specific spreading codes in a mixture of signals. The receiver-specific features analyzed in here are the front-end impairments of the receiver antenna. In particular, we focus on two distinct scenarios: real data collected under open-sky conditions from two different antennas and simulated data collected from a Spectracom simulator in the absence of noise. Additional scenarios are available on-demand (could not be uploaded due to Zenodo maximum uploading limits). The raw data generated with Spectracom simulator is provided without noise and it can thus be used to define more advanced scenarios. For example, the interested user can add different noise levels, multipath profiles, and/or interference signal components to the recorded data. In the datasets provided, we considered different signal and constellation combinations which can be used for defining complex scenarios including spoofing attacks. This can be done using dedicated software, such as Matlab. In this paper, we focus the possible use of the dataset in future RF fingerprinting applications, motivated by the fact that low-cost low-power solutions for GNSS tracking and transmitter-receiver identification are more and more on demand. Radio Frequency (RF) Fingerprinting (FP) is a relatively new concept [1–5] focusing on identifying signals based on the hardware characteristics in the communication chain. A particular case of RFF problem is the problem of identifying transmitters for more secure communications, as a modality to distinguish genuine transmitters from 'fake' or 'attacking' ones, such as spoofers and jammers. RFF concept is based on the idea that each radio transmitter, as well as each radio receiver, has unique features, not only due to the specific coding and modulation of the transmitted signals, but also due to the hardware imperfections of its various front-end blocks at transmitter and receiver sides, such as band-pass filters, local oscillators, or power amplifiers. Such features could, in theory, identify a signal, for example if coming from a spoofer or from a genuine transmitter and it could also, in theory, identify what satellite signals are present on the sky, based on the idea that each mixture of certain signals has its own 'features' or 'patterns'. Similarly with human unique fingerprints, each signal or mixture of signals has its unique features, referred to as *fingerprints*. Such fingerprints can be, for example, a combination of modulation and spreading codes of the signals in the mixture, the power amplifier non-linearities of the transmitters, the I/Q imbalances and phase noises due to the local oscillator, and various transients due to front-end filtering. The combination of these various (and typically random) effects generated by the transmitter hardware blocks creates the transmitter-specific fingerprints or features. If the fingerprints of a genuine transmitter are known (e.g., saved in a training database), then machine learning algorithms for signal classification could be applied to distinguish the genuine signals (i.e., those coming from GNSS satellites on sky) from the fake signals (e.g., those coming from ground spoofers generated with GNSS signal simulators). Of course, the wireless channel effects and the hardware characteristics of the receiver antennas should also be take into account for the best RFF results. RF FP approaches have been so far very little used in the context of GNSS [6]. Nevertheless, the ability to distinguish between genuine GNSS signals and those sent via a GNSS simulator or other GNSS fake transmitter can have a tremendous positive impact on the security of future GNSS-based applications. In order to be able to analyze the transmitter-specific and receiver antenna-specific features of GNSS signals and the GNSS signals generated via simulators, databases of genuine and simulator-based GNSS signals are needed to be made available to the research community. A characterization of our datasets and the methodology to collect them follows in Section 3. The focus of the RF FP analysis in this paper is to distinguishing real data from different antennas and also distinguishing clean data with different modulations and different number of satellites. To summarize, the purpose of this paper is, first, to offer a detailed data description of the provided dataset, and second, to show an example of how such data can be used in a limited-scope example of RFF. The main gap in the research community that we address here is the lack of open-access I/Q raw GNSS datasets that could enable researchers to study the GNSS signal characteristics in sampled domain. The RFF example that we show in here is only one possible example of how the provided data can be used, but such data opens the paths for many other possible innovative directions. Few examples of other possible uses of our data are also given in Section 5.

2. Related Work to RFF

Wireless transmitter identification based on transient signal detection has been previously studied in References [5,7–10]. The RFF based on transients rely on the assumption that there are some energy imbalances in the hardware of the transmitters, which may create random transitions (or transients) in a transmitted signal. Such transients can be used to identify a transmitter. Transient signal identification has been studied so far for various terrestrial networks, such as based on Bluetooth Low Energy (BLE) [5,9], Internet of Things (IoT) [10], Wireless Local Area Networks (WLAN) [7], or generic cognitive terrestrial radio networks [8], but it has not been studied yet in the context of GNSS, to the best of the Authors' knowledge.

Transmitter-specific features such as phase noises, I/Q imbalance and power amplifier non-linearities have been used for transmitted identification based on WLAN data in Reference [3]. Again, such hardware-specific features have not been studied so far in the GNSS context.

Research has also been done in the field of radio identifications, for example in References [11–13]. The authors of Reference [11] study the hardware impairments and investigates the Convolutional Neural Networks (CNN). The authors of Reference [12] provide a tutorial of device fingerprinting in a wireless device context, while Reference [13] introduces Permutation Entropy (PE) methods to identify devices by evaluating the level of chaos in the received signals.

3. Data Description

3.1. Laboratory Setup

The laboratory setup is mainly composed by the following equipment, as illustrated also in Figure 1:

1. Spectracom GSG-64: A multi-frequency and multi-system GNSS signal generator.
2. USRP RIO 2954R: Software-defined radio platform, two receive channels with 80 MHz/channel 70 of real-time bandwidth. It has a MXI-Express kit (PCI-based PXI controllers +x4 MXI-Express).
3. Lenovo P510 computer: Host computer (Intel Xeon CPU E3-1225 v5 @ 3.30 GHz, 32GB RAM, 256 GB SSD hard disk).
4. Tallysman TW3972 and Novatel GPS-703-GGG antennas: Triple Band GNSS Antennas placed on the roof.

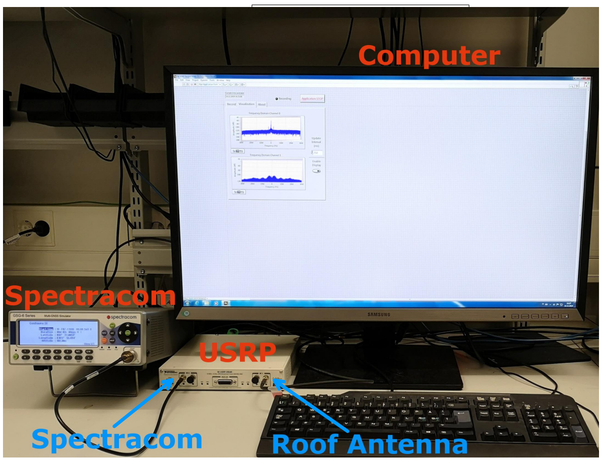


Figure 1. Laboratory setup for the signal recording.

Figure 1 shows the main equipment and the setup used for recording and generating the GNSS signals in the laboratory. The Spectracom is in charge of generating GNSS simulated data based on the set scenarios that will be presented in Section 3.3. The Universal Software Radio Peripheral (USRP) receives the Spectracom-generated data at the same time with GNSS data collected from one or both of the roof antennas (i.e., either one antenna at a time in parallel with Spectracom signal, or both roof antennas with no signal from the Spectracom). This setup allows one to collect data at the same time, with the same hardware and clock features minimizing the differences introduced by the reception platform. In addition, recording using both channels at the same time (and using the same clock) we guarantee that the same data is recorded when both roof antennas (one in each channel) are used. The only difference will be the antenna sued to capture the signal. The best-case situation is assumed during the recording, when a spoofer transmitter would be synchronized to the genuine-signal transmitter. The roof antenna and the Spectracom simulator are connected to the USRP via wired cables, namely coaxial cables with SubMiniature version A (SMA) connector of approximately 2 m and 50 cm length, respectively. The USRP acts as a front-end and down-converts the RF signals to baseband, digitizes them, and sends the complex IQ-data stream via a MXI Express link to the host PC. The host PC runs LabView software in order to control the USRP. It also receives the data streams from each channel, and it saves each of them in a flat binary file (containing binary I/Q data streams). The files can then be read and further processed.

The Spectracom GSG-64 GNSS signal generator generates the RF GNSS signals and transmit them to the USRP RIO 2954R through a RF conductor. The GNSS signal generator supports multiple signals from different satellites and constellations. The software of the Spectracom GSG-64 simulator was used directly to set up the GNSS scenario and signal parameters in order to generate the signals in the different frequency bands.

The USRP receives signal on two channels (e.g., one from roof antenna and one from the Spectracom generator), and processes each signal separately. In addition to the two signals, the USRP takes a Pulse-Per-Second (PPS) signal from the Spectracom as a clock input. This input is required to synchronize both channels so the same sample in both channels is recorded at the same time. The USRP also acts as a front-end: the two signals are down-converted in a direct down-conversion block to baseband (i.e., zero Intermediate Frequency), down-sampled via a high rate Analog-to-Digital Converter (ADC) and then the amplitude levels are quantized to obtain digital data streams. The USRP provides I/Q data streams for each channel. Each stream is an array of complex, 16-bit signed integer data (range −32,768 to +32,767). The real and imaginary components of the data correspond to the in-phase (I) and quadrature-phase (Q) data, respectively and are interleaved in the array ([I, Q, I, Q, ...]).

Besides the signal generated with the Spectracom, we also got signal from two different GNSS antennas placed in the roof: a Tallysman TW3972 and a Novatel GPS-703-GGG. Both antennas are able to receive from the three main frequency bands and from the four main constellations: GPS L1/L2/L5, GLONASS G1/G2/G3, BeiDou B1/B2 and Galileo E1/E5a+b (in addition of signals from the L-band correction services).

3.2. Measurement Parameters

The main parameters used during the recordings and signal generation are summarized in Tables 1 and 2, for the USRP and the Spectracom, respectively. Table 1 shows that the recordings Sampling Frequency was set to 50 MSamples/s. The quantization bits resolution is chosen relatively high (16 bits) due to the fact that for the further RF FP analysis, high resolution (and also a high sampling rate) data is typically preferred due to the higher amount of information that can be processed in order to extract a higher number (or more accurate) FP features. Each recording was chosen to be about 20 s duration, in order to fulfill the trade-off of having enough data to process and keeping a file size relatively low. No IF frequency was used during the recordings, all recordings were carried out directly in baseband. The USRP power amplifier gain was set to 30 dB, in order to record the signal with enough power (especially the one coming from the roof antennas, which was considerably attenuated due to propagation).

Table 1. Universal Software Radio Peripheral (USRP) parameters during recording.

Sampling Frequency	50 MSamples/s
Quantization Bits	16 bits
Recording Duration	20 s
Approximated File Size per recording	4 Gb
IF	0 Hz (baseband)
Gain	30 dB

Table 2. Spectracom signal parameters.

Transmit Power per Satellite	−70 dBm
Additive White Noise Channel	No-noise
Channel Effects	No channel effects
Simulated Receiver Movement	Static

Table 3 summarizes the main parameters for both roof antennas used during the recordings. Both antennas are quite similar, and both are compatible with the main GNSS constellations.

Table 3. Antenna specifications.

	Tallysman TW3972		**Novatel GPS-703-GGG**	
Compatible Constellations	GPS L1/L2/L5, GLONASS G1/G2/G3, BeiDou B1/B2 and Galileo E1/E5a+b			
Noise Figure	2.5 dB		2 dB	
Out of Band Rejection	L1/E1/B1/G1 <1450 MHz >30 dB >1690 MHz >30dB >1730 MHz >40 dB	L5/E5/L2/G2 <1050 MHz >45 dB <1125 MHz >30 dB >1350 MHz >45 dB	L1/E1/B1/G1 ± 100 MHz 30 dBc	L5/E5/L2/G2 ± 200 MHz 50 dBc
LNA Gain	37 dB		29 dB	
Filter Bandwidth	L1/E1/B1/G1 1525 MHz–1606 MHz	L5/E5/L2/G2 1164 MHz–1254 MHz	L1/B1/E1/G1 1551.5 MHz–1608.5 MHz	L5/E5/L2/G2 1165.5 MHz–1238.5 MHz
Dimensions	66 mm diameter × 21 mm		185 mm diameter × 69 mm	

3.3. Measurement Scenarios

Table 4 summarizes the list of different recorded scenarios. The recordings are coming from both real (antenna) and simulated (Spectracom) data. One of the constraints of the antenna recordings is that we were not able to choose which constellation and satellites we wanted to record. All the signals from all the satellites and constellations that share the same frequency bands are received at the same time. Thus, the specific amount of satellites recorded can not be pre-set, but needs to be determined after acquiring and tracking the GNSS signal collected with the roof antennas. The number and the identity of the satellites depend on the specific time the recordings were done. Since the recordings were done many times at different hours, Table 4 does not provide any specific amount of satellites present in the recorded signal. One can find out the present satellites by using a GNSS software receiver, for example the one provided in open access by Reference [14].

With the Spectracom simulated data, we were able to select the specific satellites and constellations we wanted to include in the simulated signal. With the Spectracom data we recorded individual signals from different constellations and also the combination of some of them. Therefore, we recorded signals with only a single satellite present, five present satellites and 10 present satellites. When only a single satellite was present, we considered different Pseudo-Random Number (PRN) codes. The different scenarios shown in Table 4 were recorded 10 different times, in order to be able to extract the statistical behaviour of the recordings with FP. This makes a total number of 580 Spectracom recordings, and 40 antenna recordings (20 with each antenna). So the total number of data is composed by 620 signals. For each of the Spectracom-based recordings we did not consider either channel or noise effects, as it is specified in Table 2.

Table 4. List of recorded scenarios. Channel effects and noise are not considered during the Spectracom recordings in order to have a clean reference signal. * Due to the different recording carried out at different days/times, no specific number of satellites can be given, as they vary from iteration to iteration. The number and identity of the present satellites can be found by performing the acquisition of the recorded signal using a Global Navigation Satellite System (GNSS) software receiver, for example, as in Reference [14] or a signal GNSS planner such as Trimble [15]. The Cartesian coordinates of the antennas are: X = 2,795,125.7571 m, Y = 1,236,112.0878 m Z = 5,579,645.6020 m and X = 2,795,123.0213 m, Y = 1,236,117.7508 m Z = 5,579,645.089 m, for Novatel and Tallysman, respectively. The signals were collected on 12 November 2019.

Device	Scenario ID	Constellation	Amount of Satellites
Antenna Tallysman	A1	GPS L1 + GALILEO E1 + BeiDou B1	Variable according to date and time of the recording *
	A3	GPS L5 + GALILEO E5 + BeiDou B2	Variable according to date and time of the recording *
Antenna Novatel	A2	GPS L1 + GALILEO E1 + BeiDou B1	Variable according to date and time of the recording *
	A4	GPS L5 + GALILEO E5 + BeiDou B2	Variable according to date and time of the recording *
Spectracom	S1–S9	GPS L1	1 (7 recordings with different PRN), 5 and 10
	S10–S18	GPS L5	1 (7 recordings with different PRN), 5 and 10
	S19–S27	GALILEO E1	1 (7 recordings with different PRN), 5 and 10
	S28–S36	GALILEO E5	1 (7 recordings with different PRN), 5 and 10
	S37–S45	GLONASS G1	1 (7 recordings with different PRN), 5 and 10
	S46–S54	BeiDou B1	1 (7 recordings with different PRN), 5 and 10
	S55–S56	GPS L1 + GALILEO E1 + BeiDou B1	1 and 5 per constellation
	S57–S58	GPS L1 + GALILEO E1 + BeiDou B1 + GLONASS G1	1 and 5 per constellation

4. Examples of Data Analysis

While the provided dataset can be used for multiple purposes and it is made available to the research community without any restrictions on its uses or applicability, in our studies we investigated parts of the collected datasets for the purpose of RF FP, as described in the next sections. Figure 2 shows a block diagram with the processing steps carried out in this paper for detection and classification of RF FP. First of all the signal coming from the roof antenna or the Spectracom is stored as I/Q samples. Next, a time-frequency transform of the data is applied, in order to produce an image with the results of the transform, which will be post-processed and saved for further use. After having enough of such images, an image feature extraction is performed in order to train an specific Machine Learning algorithm (in our case Support Vector Machine (SVM)), which will be the one performing the classification of the images.

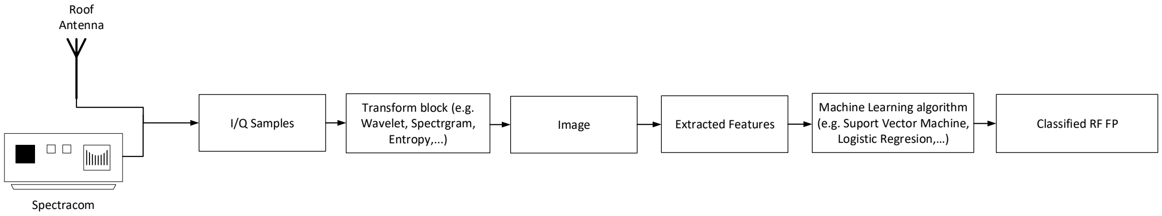

Roof Antenna
Spectracom
I/Q Samples
Transform block (e.g. Wavelet, Spectrgram, Entropy,...)
Image
Extracted Features
Machine Learning algorithm (e.g. Suport Vector Machine, Logistic Regresion,...)
Classified RF FP

Figure 2. Block diagram showing the processing steps.

4.1. *Transforms for RF FP Feature Extraction*

Several transforms can be performed before further classifications, as these transforms reveal the data behaviour in the time-frequency domain and may identify the hidden signal fingerprints. We then used these fingerprints (or features) to conduct machine learning methods in order to identify the source of the recorded signal. We remark that in order to keep the consistency with the numerical analysis in this paper, we introduce the transforms in their discretised version.

4.1.1. Wavelet Transform

The wavelet transform is a time-bounded frequency-bounded transform and gives a way of representing a signal by shifting and scaling a so-called 'mother wavelet'. The mother wavelet is defined as a finite or fast fading function (e.g., sinc function, etc.). The major strong point of wavelet transform is its multi-resolution analysis, due to its intelligently adjusting of parameters according to different frequencies.

Consider a signal $x[n]$, the Continuous Wavelet Transform (CWT) [16,17] is,

$$X(a,b) = \sum x[n]\Psi^*_{a,b}(n), \tag{1}$$

where $\Psi^*_{a,b}(n)$ is the complex conjugate of $\Psi_{a,b}(n)$, which is given by,

$$\Psi_{a,b}(n) = \frac{1}{|a|^{j/v}}\Psi\left(\frac{n-b}{a^{j/v}}\right), \tag{2}$$

where $\Psi(n)$ is the mother wavelet function, a is the scaling factor, b is the shifting parameter, $j = 1,2,3\ldots$ and $v = 2,3,4\ldots$.

The $\Psi_{a,b}(n)$ in Discrete Wavelet Transform (DWT) [18] is defined as,

$$\Psi_{a,b}(n) = \frac{1}{\sqrt{2^j}}\Psi\left(\frac{n-2^j b}{2^j}\right). \tag{3}$$

In both CWT and DWT, there exists many families of Wavelet mother functions (e.g., 'symlet' or 'haar') for different emphases of signal representations.

4.1.2. Spectrogram

The spectrogram [19,20] is a common method to analyze signals in time-frequency domain. The spectrogram is the squared value of the Short-Time Fourier transform (STFT). Again, we consider a signal $x[n]$, the STFT is in the form,

$$X(m,f) = \sum x[n]w[n-m]e^{-i2\pi fn}, \tag{4}$$

where $w[\cdot]$ is a window function, m is the time shift, f is the frequency.

The spectrogram of $x[n]$ yields to,

$$\text{Spectrogram}\{x[n]\} = |X(m,f)|^2. \tag{5}$$

4.1.3. Wigner-Ville Distribution

The Wigner-Ville distribution (WVD) is another way of time-frequency analysis and could provide high-resolution analysis of signals. For a discrete signal $x[n]$, the WVD is expressed as,

$$WVD(n,k) = \sum_{m=-N}^{N} x\left(n+\frac{m}{2}\right)x^*\left(n-\frac{m}{2}\right), \tag{6}$$

where $x^*(\cdot)$ is the complex conjugate of $x(\cdot)$.

4.1.4. Teager-Kaiser Energy Operator

The Teager-Kaiser energy operator (TKEO) could estimate the instantaneous energy of signals, for a discrete real signal $x[n]$, the TKEO is in the form of,

$$\Phi\{x[n]\} = x^2[n] - x[n+1]x[n-1]. \tag{7}$$

For a discrete complex signal $y[n]$, the TKEO is given by,

$$\Phi\{y[n]\} = y^*[n]y[n] - \frac{1}{2}\Big\{y^*[n+1]y[n-1] + y[n+1]y^*[n-1]\Big\}, \tag{8}$$

TKEO as is shown in Equation (8) is used to perform the simulations in Section 4.3.

4.2. Machine Learning for RF FP Classification

The machine learning methods are labelled with efficiency and versatility in many fields, provided the context of time series data, we mainly investigateLogistic Regression (LR) and SVM methods in the following.

4.2.1. Logistic Regression

The essence of LR is the Logistic Probability Function (LPF), defined as

$$\Pr(x) = \frac{1}{1+e^{-x}}. \tag{9}$$

The key idea is to project all the data into the interval $(0,1)$ through the LPF of (9). After this projection operation, the LR uses maximum likelihood to achieve the classifications. If we denote $\mathbf{w}$ as the projection vector, $\mathbf{x}$ as the input data, $y = \{-1,1\}$ as the classes. The loss function of LR can be given by,

$$loss(\mathbf{w}) = \sum \ln\left[1 + e^{y(\mathbf{w}^T\mathbf{x}+b)}\right] + C\mathbf{w}^T\mathbf{w}, \tag{10}$$

where b is the introduced bias, C is the penalty parameter.

4.2.2. Support Vector Machine

In a two-class scenario, SVM classifies classes by maximising the margin width between classes. The SVM is flexible that can be adapted in high-dimension and/or non-linear space, due to changeable kernels in SVM. A brief review of 'kernel tricks' is addressed below.

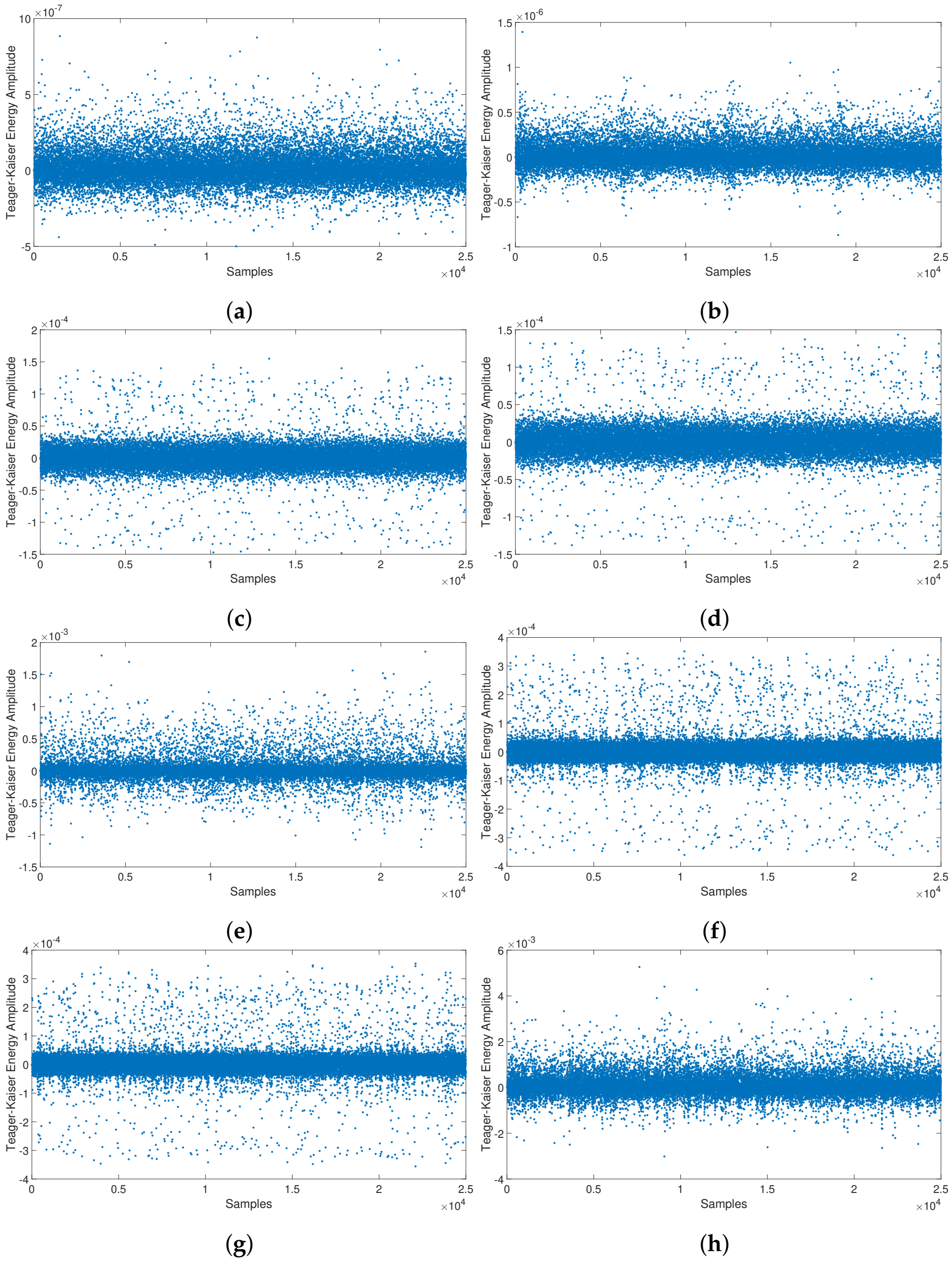
Consider a two-dimension space, the kernel is marked as $k(\mathbf{x},\mathbf{y})$.

- linear kernel: $k(\mathbf{x},\mathbf{y}) = \mathbf{x}\cdot\mathbf{y}$;
- polynomial kernel: $k(\mathbf{x},\mathbf{y}) = \big(\mathbf{x}\cdot\mathbf{y}\big)^d$, d is the order;
- Gaussian kernel (or 'rbf'): $k(\mathbf{x},\mathbf{y}) = \exp\Big(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\Big)$;
- sigmoid kernel: $k(\mathbf{x},\mathbf{y}) = \tanh(a\mathbf{x}\cdot\mathbf{y}+b)$, $a > 0$ and $b < 0$.

In this work, we employed Gaussian kernel, as this is the most common kernel used in literature with SVM classifiers. Nevertheless, our data provided in open-access can be further studied with various other kernels, as described above.

4.3. Results

Figure 3 shows snapshot examples of TKEO plot for each considered scenario from Table 4, where the Teager-Kaiser (TK) energy is depicted for each sample. Small differences between different scenarios are visible by eye, but it is to be emphasized that the images shown in Figure 3 are only snapshot images, while the machine learning classifier will use a large amount of such images to be able to identify a certain scenario.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**) (**g**) (**h**)

Figure 3. Examples of TKEO images for each considered scenario used for classification with SVM. Image resolution is 256×256 pixels. (**a**) Scenario A1; (**b**) Scenario A2; (**c**) Scenario S1; (**d**) Scenario S2; (**e**) Scenario S9; (**f**) Scenario S19; (**g**) Scenario S20; (**h**) Scenario S27.

Figures 4 and 5 show the confusion matrix results after applying the classification SVM, algorithm described in Section 4.2.2, for real and simulated data, respectively. The classification problem addressed in Figure 4 is to identify between two front-end receivers, namely one using a Novatel antenna and another one using a Tallysman antenna, when signals received from both antennas are synchronized and the antenna baseline was about 6.3 m (the Cartesian coordinates of the antennas are: X = 2,795,125.7571 m,Y = 1,236,112.0878 m Z = 5,579,645.6020 m and X = 2,795,123.0213 m,Y = 1,236,117.7508 m Z = 5,579,645.089 m, for Novatel and Tallysman, respectively). The classification problem addressed in Figure 5 is to identify between six classes of GNSS signal mixtures. Each class has a different number of satellites and/or system type (i.e., GPS or Galileo).

The confusion matrix shows how accurate a classifier is, in terms of how well it classifies and miss-classifies the test data set into the different classes it has. In our examples, we have split the simulations into two parts, namely a classification based on two classes with roof-antenna signals

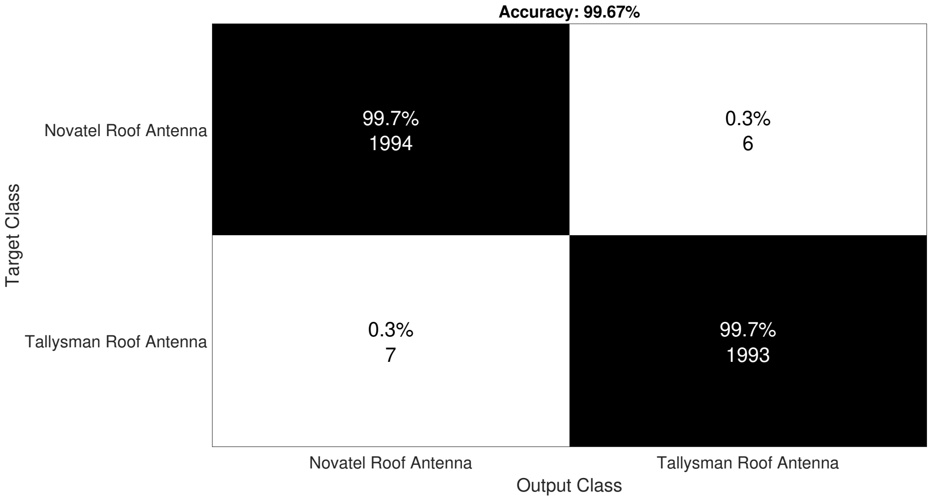
(Figure 4) and a classification based on six classes with Spectracom-generated signals (Figure 5). This was done in order to can compare the results fairly in similar conditions (e.g., noise, power, etc.).

Figure 4 shows that the average accuracy for the classifier when only two classes exists is more than 99%.

Regarding Spectracom-generated data, the confusion matrix-results are depicted in Figure 5. In this case, the average accuracy is lower than using real/roof-antenna data, but it is still more than 92%. Scenarios with 10 different satellites can be perfectly classified (100% accuracy). It means that the classifier can differentiate between signals with a single satellite and several satellites given the same signal structure (namely Global Positioning System (GPS) or Galileo).

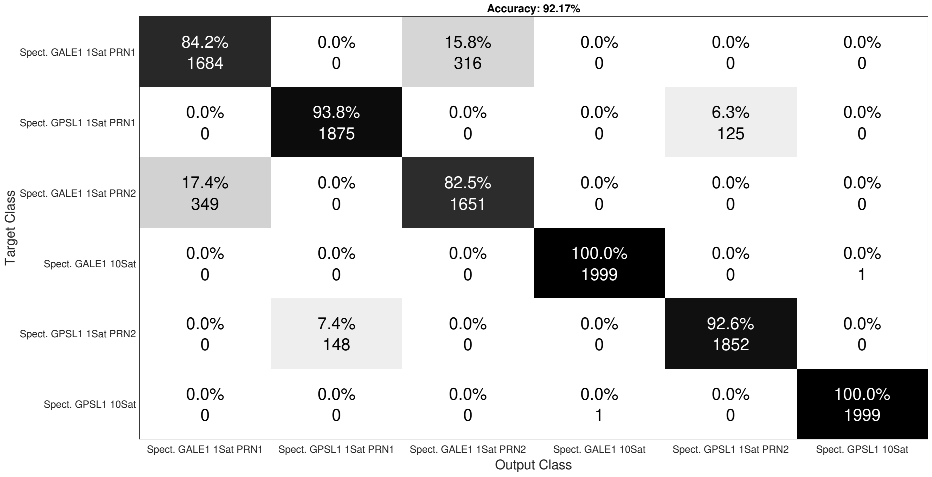
In addition, we observe in Figure 5 that when only one satellite is present in the signal, the classifier is able to determine to which PRN it belongs, although with some mis-classification error. For example, if we try to differentiate the scenarios with Galileo signal and PRN's 23 and 2, the detector miss-classifies about 17% of cases. Also, with GPS and PRN's 30 and 15, the detector only miss-classifies less than 10% of cases. As we observe, the different signal shape of Galileo and GPS affects the accuracy of the detector.

Such results are promising in the context of low-cost acquisition of signals too (e.g., if one could delegate to the cloud the identification of satellites on the sky based on the features of various mixtures of satellite signals, this would reduce the computational burden at the receiver side).



	Novatel Roof Antenna	Tallysman Roof Antenna
Novatel Roof Antenna	99.7% 1994	0.3% 6
Tallysman Roof Antenna	0.3% 7	99.7% 1993

Figure 4. Confusion matrix results from SVM using TKEO with real data.



	Spect. GALE1 1Sat PRN1	Spect. GPSL1 1Sat PRN1	Spect. GALE1 1Sat PRN2	Spect. GALE1 10Sat	Spect. GPSL1 1Sat PRN2	Spect. GPSL1 10Sat
Spect. GALE1 1Sat PRN1	84.2% 1684	0.0% 0	15.8% 316	0.0% 0	0.0% 0	0.0% 0
Spect. GPSL1 1Sat PRN1	0.0% 0	93.8% 1875	0.0% 0	0.0% 0	6.3% 125	0.0% 0
Spect. GALE1 1Sat PRN2	17.4% 349	0.0% 0	82.5% 1651	0.0% 0	0.0% 0	0.0% 0
Spect. GALE1 10Sat	0.0% 0	0.0% 0	0.0% 0	100.0% 1999	0.0% 0	0.0% 1
Spect. GPSL1 1Sat PRN2	0.0% 0	7.4% 148	0.0% 0	0.0% 0	92.6% 1852	0.0% 0
Spect. GPSL1 10Sat	0.0% 0	0.0% 0	0.0% 0	0.0% 1	0.0% 0	100.0% 1999

Figure 5. Confusion matrix results from SVM using TKEO with Spectracom data.

5. Conclusions and Open Directions

We have provided in open-access a multi- purpose dataset of raw GNSS measurements from sky satellites and Spectracom simulator. We have shown one particular use case of such data, namely how our data can be used for RF fingerprinting problem. As an example, a combination of machine learning algorithms and Teager-Kaiser transform was used to identify the signal type based on our collected datasets. We showed that, TKEO is useful to identify the signal features, both at the transmitter side (mixture of signals) and at the receiver side (antenna type). In our results, we could classify the receiver antenna type with an average accuracy higher than 99%. We have also shown that such machine learning principle combined with TKEO can separate between different GNSS constellations and satellite numbers, with an average accuracy of 92% in the considered scenarios. The RFF example shown here is only an illustrative example of how the provided raw data can be used in research. Nevertheless, the provided datasets can serve multiple purposes besides the transmitter type identification, such as GNSS time-frequency characterization and feature extraction through various transforms, signal acquisition and tracking studies with Spectracom-generated and roof-antenna-collected data, GNSS data compression studies for example for future low-cost low-power navigation solutions and cloud GNSS processing, as well as development of spoofing identification mitigation algorithms. As the I/Q datasets are huge and we provide high-resolution data collected at high sampling frequencies, an important open challenge remains to find suitable transforms and compression methods to store relevant GNSS transmitter features with small datasizes.

6. Dataset Repository And License

The recorded signals used for achieve this results as well as the script to read the signals are available in open access at Zenodo [21] in open access with CC4 license.

Author Contributions: R.M.F.: Software development, hardware setup, conceptualization, running simulations, writing–original draft preparation, data analysis; W.W.: Conceptualization, Writing—Original draft preparation, data analysis; A.S.A.: Hardware setup, measurements, running simulations, Writing—Review and editing; E.S.L.: methodology, Writing—Review and editing, supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by Tampere University Rector's grant for PhD studies. Additional support has been obtained from the Academy of Finland project ULTRA (number 328226).

Acknowledgments: The authors would also like to express their warm thanks to the anonymous reviewers for their constructive comments.

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

Abbreviations

ADC	Analog-to-Digital Converter
BLE	Bluetooth Low Energy
CNN	Convolutional Neural Networks
CWT	Continuous Wavelet Transform
DWT	Discrete Wavelet Transform
FP	Fingerprinting
GPS	Global Positioning System
GNSS	Global Navigation Satellite Systems
IF	Intermediate Frequency
IoT	Internet of Things
LR	Logistic Regression
LPF	Logistic Probability Function
PE	Permutation Entropy
PPS	Pulse-Per-Second

PRN	Pseudo-Random Number
RF	Radio Frequency
RFF	Radio Frequency Fingerprinting
SDR	Software Defined Radio
SMA	SubMiniature version A
STFT	Short-Time Fourier transform
SVM	Support Vector Machine
TK	Teager-Kaiser
TKEO	Teager-Kaiser energy operator
USRP	Universal Software Radio Peripheral
WLAN	Wireless Local Area Networks
WVD	Wigner-Ville distribution

References

1. Bertoncini, C.; Rudd, K.; Nousain, B.; Hinders, M. Wavelet Fingerprinting of Radio-Frequency Identification (RFID) Tags. *IEEE Trans. Ind. Electron.* **2012**, *59*, 4843–4850. [CrossRef]
2. Patel, H.J.; Temple, M.A.; Baldwin, R.O. Improving ZigBee Device Network Authentication Using Ensemble Decision Tree Classifiers With Radio Frequency Distinct Native Attribute Fingerprinting. *IEEE Trans. Reliab.* **2015**, *64*, 221–233. [CrossRef]
3. Sankhe, K.; Belgiovine, M.; Zhou, F.; Angioloni, L.; Restuccia, F.; D'Oro, S.; Melodia, T.; Ioannidis, S.; Chowdhury, K. No Radio Left Behind: Radio Fingerprinting Through Deep Learning of Physical-Layer Hardware Impairments. *IEEE Trans. Cogn. Commun. Netw.* **2019**, 1. [CrossRef]
4. Aghnaiya, A.; Ali, A.M.; Kara, A. Variational Mode Decomposition-Based Radio Frequency Fingerprinting of Bluetooth Devices. *IEEE Access* **2019**, *7*, 144054–144058. [CrossRef]
5. Ali, A.M.; Uzundurukan, E.; Kara, A. Assessment of Features and Classifiers for Bluetooth RF Fingerprinting. *IEEE Access* **2019**, *7*, 50524–50535. [CrossRef]
6. Borio, D.; Gioia, C.; Cano Pons, E.; Baldini, G. GNSS Receiver Identification Using Clock-Derived Metrics. *Sensors* **2017**, *17*, 2120. [CrossRef] [PubMed]
7. Shi, Z.; Liu, M.; Huang, L. Transient-based identification of 802.11b wireless device. In Proceedings of the 2011 International Conference on Wireless Communications and Signal Processing (WCSP), Nanjing, China, 9–11 November 2011; pp. 1–5, doi:10.1109/WCSP.2011.6096906. [CrossRef]
8. Hu, N.; Yao, Y. Identification of legacy radios in a cognitive radio network using a radio frequency fingerprinting based method. In Proceedings of the 2012 IEEE International Conference on Communications (ICC), Ottawa, ON, Canada, 10–15 June 2012; pp. 1597–1602, doi:10.1109/ICC.2012.6364436. [CrossRef]
9. Ali, A.M.; Uzundurukan, E.; Kara, A. Improvements on transient signal detection for RF fingerprinting. In Proceedings of the 25th Signal Processing and Communications Applications Conference (SIU), Antalya, Turkey, 15–18 May 2017; pp. 1–4, doi:10.1109/SIU.2017.7960417. [CrossRef]
10. Köse, M.; Taşcioğlu, S.; Telatar, Z. RF Fingerprinting of IoT Devices Based on Transient Energy Spectrum. *IEEE Access* **2019**, *7*, 18715–18726. [CrossRef]
11. Riyaz, S.; Sankhe, K.; Ioannidis, S.; Chowdhury, K. Deep learning convolutional neural networks for radio identification. *IEEE Commun. Mag.* **2018**, *56*, 146–152. [CrossRef]
12. Xu, Q.; Zheng, R.; Saad, W.; Han, Z. Device fingerprinting in wireless networks: Challenges and opportunities. *IEEE Commun. Surv. Tutor.* **2015**, *18*, 94–104. [CrossRef]
13. Deng, S.; Huang, Z.; Wang, X.; Huang, G. Radio frequency fingerprint extraction based on multidimension permutation entropy. *Int. J. Antennas Propag.* **2017**, *2017*, 6. [CrossRef]
14. An Open Source Global Navigation Satellite Systems Software-Defined Receiver. Available online: https://gnss-sdr.org/ (accessed on 19 December 2019).
15. TRIMBLE GNSS Planning Online. Available online: https://www.gnssplanning.com/ (accessed on 19 December 2019).
16. Lilly, J.M.; Olhede, S.C. Generalized Morse wavelets as a superfamily of analytic wavelets. *IEEE Trans. Signal Process.* **2012**, *60*, 6036–6041. [CrossRef]

17. Lilly, J.M. Element analysis: A wavelet-based method for analysing time-localized events in noisy time series. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **2017**, *473*, 20160776. [CrossRef] [PubMed]
18. Mallat, S.G. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **1989**, *7*, 674–693. [CrossRef]
19. Mitra, S.K.; Kuo, Y. *Digital Signal Processing: A Computer-Based Approach*; McGraw-Hill: New York, NY, USA, 2006; Volume 2.
20. Oppenheim, A.V. *Discrete-Time Signal Processing*; Pearson Education India: Bengaluru, India, 1999.
21. Ferre, R.M.; Wang, W.; Lohan, E.S. Identifying GNSS transmitters based on their RadioFrequency (RF) features—A dataset with GNSS roofantenna and Spectracom-based GNSS signals. *Zenodo* **2020**, in press. doi:10.5281/zenodo.3629290.