Cyber-threats Analytics for Detection of GNSS Spoofing

Silvio Semanjski*, **
* Department of Communication, Information, Systems & Sensors
** Royal Military Academy
Brussels, Belgium
e-mail: silvio.semanjski@rma.ac.be

Ivana Semanjski*, **
* Department of Industrial Systems Engineering and Product Design, Ghent University,
Ghent, Belgium
** Industrial Systems Engineering (ISyE), Flanders Make
Ghent, Belgium
e-mail: ivana.semanjski@ugent.be

Wim De Wilde, **
** Septentrio N.V.
Leuven, Belgium
e-mail: wim.dewilde@septentrio.com

Alain Muls*, **
* Department of Communication, Information, Systems & Sensors
** Royal Military Academy
Brussels, Belgium
e-mail: alain.muls@rma.ac.be

***Abstract*—Spoofing of the Global Navigation Satellite System (GNSS) open service (unencrypted) signal is of continuous interest to professionals and non-professional users. The main reason for this is the risk of unaware use of manipulated GNSS data, which becomes extremely relevant in all Safety-of-Life (SOL) Position-Navigation-Timing (PNT) applications, such as aircraft navigation or high precision time synchronization of traffic control systems. In this paper, we aim to develop an approach to detect spoofing of the GNSS signal based on the machine learning technique. The developed approach shows high potential in detecting the spoofed signal in the sequence of the non-spoofed GNSS signals by achieving the success rate of 96%.**

Keywords-Global Navigation Satellite System; Spoofing; Support Vector Machines; Safety-of-Life; Position-Navigation-Timing; GPS; GNSS; PNT; SVM; SOL

I. INTRODUCTION

Spoofing of the GNSS open service (unencrypted) signal is of continuous interest to both GNSS industry and users [1] due to risk of unaware use of manipulated GNSS data in Safety-of-Life PNT applications, such as aircraft navigation or high precision time synchronization of traffic control systems. With advances in digital signal processing and availability of the electronic components required to build transmit capable Software Defined Radio (SDR) type of spoofers, the threat of GNSS signal spoofing proliferates and requires effort to implement spoofing detection at the GNSS receiver level.

The GNSS measurements performed by the user's receiver contains a number of observables whose monitoring and cross-correlation can be used to detect the GNSS spoofing, latest at the stage of generating Position-Velocity-Time (PVT) solution within the receiver. One of the known spoofing techniques, the Time Synchronization Attack (TSA) [2] is based on the manipulation of GNSS receiver clock offset by exploiting clock drift (time derivative of the clock offset) estimates, affecting pseudorange measurements and consequentially PVT solution.

In this paper, we examine the potential to detect the GNSS signal spoofing by applying the machine learning approach, namely the Support Vector Machines (SVM) classification. Among several GNSS spoofing detection methods being discussed in details in [1], detection by observing time manipulation or discrepancies within the GNSS receiver proves to be a challenge [2], and its implementation requires subtle approach when compared to others (such as signal angle-of-arrival, strength, doppler shift as the relative speed between satellite/spoofer and receiver, signal-to-noise ratio, and signal polarization [3]). The approach to monitor the clock bias by employing SVM classification of multiple variables used in different processing stages within the GNSS receiver has been chosen due to dynamic characteristics of the target receivers (moving aircraft relative to the spoofer), and computational effectiveness of the algorithm expressed as a scalable runtime in regard to the number of input samples. The literature suggests that in the latter case, the SVM classification emerges as an concurrent choice [4].

The paper is composed as follows: Section 2 gives a detail insight into the data set and the method description. This is followed with the results and the discussion sections. Section 5 presents the conclusion remarks.

II. DATA AND METHOD

A. Dataset description

Spoofing dataset (with matched power attack) has been generated with a modified Spirent GNSS signal and constellation simulator connected to a Wave Field Synthesis (WFS) anechoic chamber at the Fraunhofer FORTE facility [5][6]. The six channels represent the "authentic" GNSS signal. In parallel, the six other channels have exactly the same parameters, including the simulated spoofing attack. The spoofer only gets enabled for three minutes in the test

 ISBN: 978-1-61208-681-1

scenario, so dataset includes non-spoofed and spoofed epochs. In our spoofing scenario, same used in [3], the spoofing attack hijacked the Pulse-Per-Second (PPS) output of the receiver because of the programmed clock divergence. Spoofing attack generated was an intermediate timing attack with 5 ns/s rate of time pulling.

The GNSS open services in general, such as Global Positioning System (GPS), have their navigation message modulated together with Coarse/Acquisition (C/A) code onto a carrier at the L1 frequency. The navigation message together with C/A ranging code provides users with the necessary information to generate the PVT solution. The navigation message data includes: the ephemeris parameters, required to compute the satellite coordinates, the timing parameters and clock corrections, used to compute satellite clock offset, the service parameters with satellite health information, ionospheric parameters model required to compensate for ionospheric propagation delay, and the almanac, allowing the computation complete satellite constellation required to perform rough initial localisation of the user's receiver during signal acquisition phase.

GNSS receiver decodes the navigation messages and together with use of C/A ranging codes provide observables, of which the following parameters are used in our model (Table I.).

TABLE I. OBSERVABLES OUTPUT OF THE BASEBAND PROCESSING STAGE OF GNSS RECEIVER

Parameter	**Unit**	**Description**
Receiver clock drift	ppm	Receiver clock drift relative to system time (relative frequency error)
Receiver clock bias	msec	Receiver clock bias relative to system time
Code variance	cm^2	Estimated code tracking noise variance
Carrier variance	$mcycle^2$	Estimated carrier tracking noise variance
C/N_0	dB-Hz	Carrier-to-noise density ratio per channel
PR	m	Pseudorange user-to-satellite
L	cycles	Full carrier phase

Next to the variables present in Table I., we manually labelled the records that represent spoofed signal. Hence, we created an additional variable called *Class* that indicates weather the exact record belongs to the spoofed period or not. This variable will be used as the dependent categorical variable in our classification problem.

B. Support vector machines classification

Support Vector Machines (SVM) are supervised machine learning algorithms which can be used both for classification [7][8] or regression analysis [9][10]. In further description, we will focus on the SVM classification analysis as the GNSS spoofing problem, having the categorical dependent variable with two possible values (signal spoofed or not-spoofed), corresponds to the classification problem.

The SVM classification method relies on the concept of decision hyperplanes that define decision boundaries (separate between a set of objects having different class memberships). However, in practical applications, this task is not very simple and use of structures more complex than linear ones is needed to correctly classify the objects. For this purpose different mathematical functions, also called kernels, can be used in order to map objects in the *n* dimensional space [11][12]. Such mapped objects aim to have structures that are easier to separate, based on the class membership, than the original set of objects for which the mapping was not preformed. To do so, we firstly divide our dataset in two parts: the training (Z_1) and the test dataset (Z_2). This division is made based on the 75%-25% principle, randomly sorting the 75% of data into the training set and 25% into the test set. As we wanted to obtain scalable runtime in regard to the number of input samples we selected the C-SVM classification type for our problem. The literature suggests that in such cases the C-SVM is a better option over, for example, nu-SVM classification [4]. For the applied C-SVM type, the minimization error function is defined as:

$$\frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i \quad (1)$$

subject to the constraints:

$$y_i(w^T\,\phi(x_i) + b) \geq 1 - \xi_i \quad (2)$$

$$\xi_i \geq 0 \quad (3)$$

where:

$i = 1, \ldots, N$,

w - the vector of coefficients;

C - the capacity constant;

b – constant;

ξ_i - parameters for handling non-separable data (inputs).

The index *i* labels the *N* training cases ($y \in \pm 1$ represents the class labels and x_i represents the independent variables). The ϕ stands for kernel function, which in our case is the Radial Basis Function (RBF) that transforms input to the feature space:

$$K(X_i, X_j) = \phi(X_i)\cdot\phi(X_j) = \exp\left(-\gamma\left|X_i - X_j\right|^2\right) \quad (4)$$

To map the multiclass problem into binary classification problem, we applied one-against-all approach. However, the values of capacity constants C (1) and γ (4) are important to keep the training error small and in order to generalize well [13]. Since it is not possible to know beforehand the best values of these constrains for a given problem, we applied the incremental grid-search on C, in range from 1 to 10, with the step equal to 1, and γ, in range from 0 to 0.5, with the step equal to 0.01. The values that achieved the best average 10-fold cross-validation accuracy were chosen for use on the test data. These values were 10 for C and 0.125 for γ.

For the v-fold cross-validation, the total number of cases was divided into v , where $v = 10$, sub samples $Z_1, Z_2, \ldots, Z_v$ of equal sizes ($N_1, N_2, \ldots, N_v$). The v-fold cross-validation estimate is the proportion of cases in the subsample Z that are misclassified by the classifier constructed from the subsample $Z - Z_v$. This estimate is calculated in the following way:

$$R(d^{(v)}) = \frac{1}{N_v}\sum_{(x_n, j_n)\in Z_v} X(d^{(v)}(x_n) \neq j_n) \quad (5)$$

where $d^{(v)}(x)$ is the classifier calculated from the sub sample $Z - Z_v$ and X is the indicator function for which is valid:

$X = 1$, if the statement $X\left(d^{(v)}\right) \neq j_n)$ is true

$X = 0$, if the statement $X\left(d^{(v)}\right) \neq j_n)$ is false.

The test sample estimate is the proportion of cases in the test dataset that are misclassified by the classifier constructed from the learning dataset. This estimate is computed in the following way:

$$R(d) = \frac{1}{N_2} \sum_{(x_n, j_n) \in Z_2} X(d(x_n) \neq j_n) \quad (6)$$

III. RESULTS

The overall success rate of the proposed approach was 96.4%, whereas the cross-validation error was slightly higher (96.8%). In total, 57 support vectors were used, of which 28 belonged among the "authentic" GNSS observations and 29 among the spoofed GNSS signal observations. In Table II. the overall summary of the obtained results is shown.

TABLE II. MODEL SUMMARY

	Value
Number of independents	8
SVM type	Classification type 1
Kernel type	Radial Basis Function
Number of SVs	57 (52 bounded)
Number of SVs (authentic GNSS signal)	28
Number of SVs (spoofed GNSS signal)	29
Cross -validation accuracy	96.765 %
Class accuracy (training dataset)	96.765 %
Class accuracy (test dataset)	95.359 %
Class accuracy (overall)	96.414 %

Considering the confusion matrix shown in TABLE III., none of the authentic GNSS signal records was confused to be the spoofed record. However, 3.6% of the records were misclassified as the authentic GNSS signal record, whereas they belonged among the spoofed signal records.

TABLE III. CONFUSION MATRIX

	Authentic GNSS signal	Spoofed GNSS signal
Authentic GNSS signal	88.34%	0.00%
Spoofed GNSS signal	3.64%	8.02%

The correlation matrix (Table IV.) shows the correlations among all the variables used in our model. Those marked with blue colour are significant at $p < 0.05$. One can see that this includes all the variables except the *Code variance*, for which the correlation with the *Class* indication is not statistically significant at $p < 0.05000$. The highest correlation is noted for the *Carrier variance*, indicating that 56.25% of the variations among the *Class* variable can explained by the *Carrier variance*. However, among the predictor variables, the highest correlation (0.99) is noted among the *Receiver clock bias* and *Receiver clock drift*.

Fig. 1 shows the *Receiver clock drift* per each second (on the *x* axis). One can clearly notice the indication of the spoofing period starting around 130 seconds into the test, lasting a bit less than 3 min (recording being started 50 seconds into the test). Fig. 2 shows the same period per each of the six channels (each channel corresponding to the satellite tracked). The initial jump in the *Carrier-to-Noise density ratio* clearly marks the beginning of the spoofing period, whereas the jump gets smoother after the initial jump in the value.

TABLE IV. CORRELATIONS TABLE

	Code variance [cm^2]	**Carrier variance [mcycle^2]**	**C/N$_0$ [dB-Hz]**	**PR [m]**	**L [cycles]**	**Receiver clock bias [ms]**	**Receiver clock drift [ppm]**	**Class**
Code variance [cm^2]	1.000	0.322	-0.309	0.215	0.215	0.209	-0.066	0.049
Carrier variance [mcycle^2]	0.322	1.000	0.278	0.406	0.406	0.405	0.359	0.750
C/N$_0$ [dB-Hz]	-0.309	0.278	1.000	-0.134	-0.134	-0.125	0.427	0.537
PR [m]	0.215	0.406	-0.134	1.000	1.000	0.999	0.259	0.376
L [cycles]	0.215	0.406	-0.134	1.000	1.000	0.999	0.259	0.376
Receiver clock bias [ms]	0.209	0.405	-0.125	0.999	0.999	1.000	0.257	0.375
Receiver clock drift [ppm]	-0.066	0.359	0.427	0.259	0.259	0.257	1.000	0.556
Class	0.049	0.750	0.537	0.376	0.376	0.375	0.556	1.000

 ISBN: 978-1-61208-681-1

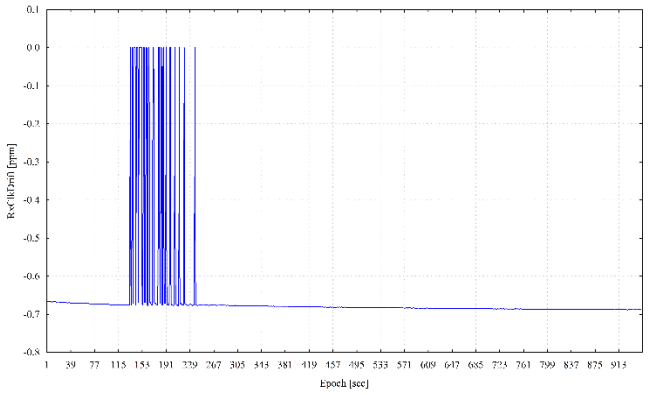

Figure 1. Receiver clock drift.

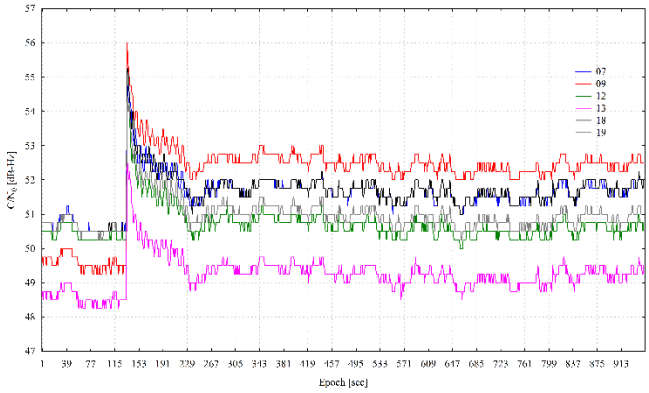

Figure 2. Carrier-to-Noise density ratio of the satellites used.

IV. DISCUSSION

In our spoofing scenario, the spoofing attack hijacked the Pulse-Per-Second output of the receiver through the programmed clock divergence. Spoofing attack generated was an intermediate timing attack with 5 ns/s rate of time pulling. The results indicate that the proposed approach can be successfully used to detect GNSS spoofing that corresponds to the above-described scenario. Due to risk of unaware use of manipulated GNSS data, this is highly relevant for all GNSS applications. However, it becomes particularly relevant when it comes to Safety-of-Life PNT applications, such as aircraft navigation, or high precision time synchronization of traffic control systems. The former proves as a challenge due to GNSS receiver being a moving target, as opposed to the latter where a target is a fixed GNSS timing receiver.

The confusion matrix indicates that the proposed machine learning based approach was able to correctly detect weather the signal spoofed or the authentic one in 96.4% of the cases. However, the remaining 3.6% of the cases were confusions made between spoofed GNSS signal records that were misclassified as the authentic ones. Considering the Safety-of-Life applications, this is less preferred scenario (over relaxed one) compared to the possibility to misclassify authentic signal as the spoofed one (over causes one). Hence, there is still a room to improve the proposed approach.

The correlation matrix indicates the statistically relevant correlation among the variables selected for our model (significant at $p < 0.05$). One can also notice very high correlation among *Receiver clock bias*, *Pseudorange*, and *Phase cycle* variables. Such a high correlation indicates that sub selection of variables would be able to explain the variation between the indication of the spoofed and the authentic GNSS signal in an equally efficient manner. Hence, our future research will focus on simplification of the model in terms of the possibility to exclude some of the considered predictor variables and their replacement with potential predictors that could affect the confusion. Although the achieved success rate is quite high, for Safety-of-Life applications, we would aim look for a model, even with the similar success rate (if not possible to achieve the highest success rate), that would result in an over causes scenario, rather than the over relaxed one.

V. CONCLUSION AND FUTURE WORK

Our research included synthetically generated GNNS signal over six channels (for six satellites) with simulation of the spoofing attack. By indicating the spoofing attack among the correct records (ones corresponding to the authentic signal), we have created a dataset that could be used for learning to recognise the spoofing attack in the machine learning approach. For the following, we have adopted the support vector machines-based approach. The achieved results show a high success rate in detecting whether the signal was spoofed or not (96.4%). However, the confusion matrix indicates that there is a space for the improvements in order to be able to use the suggested approach in the Safety-of-Life applications, such as aircraft navigation high precision time synchronization of traffic control systems. The correlation matrix also indicates that there is a possibility to improve the suggested approach, without increasing the complexity of the problem. This can be done by replacing the part of the predictor variables (those with the high correlation among them) with ones that could explain the part of the variation, among the spoofed or not-spoofed signal indication, which is not explained by the variables already present in the model.

ACKNOWLEDGMENT

The authors wish to thank Septentrio N.V. for supporting this work by providing datasets for the analyses.

REFERENCES

[1] D. Schmidt, K. Radke, S. Camtepe, E. Foo, and M. Ren, “A Survey and Analysis of the GNSS Spoofing Threat and Countermeasures,” ACM Comput. Surv., 2016.

[2] N. O. Tippenhauer, C. Pöpper, K. B. Rasmussen, and S. Capkun, “On the requirements for successful GPS spoofing attacks,” in Proceedings of the 18th ACM conference on Computer and communications security - CCS ’11, 2011.

[3] W. De Wilde et al., “Authentication by Polarization: A Powerful Anti-Spoofing Method,” in Proceedings of the 31st International Technical Meeting of the Satellite Division of the Institute of Navigation (ION GNSS+ 2018), 2018, pp. 3643–3658.

[4] C.-C. Chang and C.-J. Lin, “Training v -Support Vector

 ISBN: 978-1-61208-681-1

Classifiers: Theory and Algorithms," Neural Comput., vol. 13, no. 9, pp. 2119–2147, 2001.

[5] A. Rügamer, G. Del Galdo, J. Mahr, G. Rohmer, G. Siegert, and M. Landmann, "Testing using Wave-Field Synthesis," in Proceedings of the 26th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS+ 2013), 2013, pp. 1931–1943.

[6] C. Schirmer et al., "3D wave-field synthesis for testing of radio devices," in 8th European Conference on Antennas and Propagation, EuCAP 2014, 2014, pp. 3394–3398.

[7] S. Joo, C. Oh, E. Jeong, and G. Lee, "Categorizing bicycling environments using GPS-based public bicycle speed data," Transp. Res. Part C Emerg. Technol., vol. 56, pp. 239–250, Jul. 2015.

[8] I. Semanjski and S. Gautama, "Crowdsourcing mobility insights – Reflection of attitude based segments on high resolution mobility behaviour data," Transp. Res. Part C Emerg. Technol., vol. 71, 2016.

[9] E. I. Vlahogianni, "Optimization of traffic forecasting: Intelligent surrogate modeling," Transp. Res. Part C Emerg. Technol., vol. 55, pp. 14–23, Jun. 2015.

[10] J. Wang and Q. Shi, "Short-term traffic speed forecasting hybrid model based on Chaos – Wavelet Analysis-Support Vector Machine theory," Transp. Res. Part C, vol. 27, pp. 219–232, 2013.

[11] B. Schlkopf and A. J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning). Cambridge: The MIT Press, 2001.

[12] L. H. Hamel, Knowledge Discovery with Support Vector Machines. Hoboken: Wiley-Interscience, 2011.

[13] D. Anguita and L. Oneto, "In – sample Model Selection for Support Vector Machines," in The 2011 International Joint Conference on Neural Networks, 2011.

 ISBN: 978-1-61208-681-1