Deep Learning in RAT and Modulation Classification with a New Radio Signals Dataset

Ingrid Nascimento, Flavio Mendes, Marcus Dias, Andrey Silva and Aldebaro Klautau

Abstract— **The automatic classification (or identification) of modulation schemes and radio access technologies (RAT) find several applications in military and cognitive radio systems. As in other domains, deep learning has been applied to classification problems in telecommunications. As other machine learning approaches, assessing deep learning depends on the available datasets. However, the evaluation of previous work in modulation classification was done only with simulated signals, which may not properly represent realistic scenarios. In this paper, we revisit modulation classification schemes and also conduct experiments in RAT classification. One of the contributions is a new public dataset of digitized signals with LTE and GSM signals, both simulated and digitized. We then compare deep learning with other classifiers and observe that with a more comprehensive set of features than used in recent works, deep convolutional networks do not significanytly outperform other classifiers under the tested conditions. The results also allow to draw conclusions regarding the performance of classifiers under mismatched training and test sets, such as training only with simulated signals and testing with digitized waveforms obtained from commercial mobile networks.**

Keywords— **Deep learning, Radio Access Technology Classification, Automatic Modulation Classification.**

I. Introduction

Given an unknown received transmission format, the automatic modulation classification (AMC) is the process of determining which modulation the received waveform uses. This topic receives great attention in military applications as eletronic warfare and spectrum surveillance to identify threats in transmitted signals. Futher, this technique is applied in cognitive radio systems for spectrum awareness to avoid interference between users [1], besides data transmission optimization and spectrum allocation improvements [2]. A related approach is the Radio Access Technology (RAT) which differentiates between two or more communication standards. One of the employments of RAT recognition is in cognitive radio for spectrum sensing, which aim to identify licensed primary users in a specific spectrum band to avoid interference with secundary users [3]. RAT classification was also applied to detect different types of wireless system with framework based on maximum likelihood estimation [4]. Other studies as in [5] presents a classifier that differentiate between LTE and WiMax technology by the analysis of cyclostationary features. A relevant previous work for the scope of this paper is [6], in which the authors use a statistical test based on the probability distributions to distinguish LTE and GSM.

The authors are with LASSE - 5G & IoT Research Group, Federal University of Pará (UFPA), Belém-PA, CEP 66075-110, Brazil, E-mails: {ingrid.nascimento, flavio.mendes, marcus.dias, andrey.silva}@itec.ufpa.br and aldebaro@ufpa.br

Deep Learning applied to AMC is a recent trend due to its great performance in different types of data in comparison with traditional machine learning techniques. For example, a deep convolutional neural (CNN) model was used in [7] for the classification of eight digital modulations (BPSK, QPSK, 8PSK, 16QAM, 64QAM, BFSK, CPFSK and 4 PAM) and two analog modulations (WB-FM and AM-DSB). In [8], the author extends the work done in [7] and shows that the Convolutional Long Short-term Deep Neural Network (CLDNN) architecture gives better results than the CNN adopted in [7].

In machine learning, as important as the learning algorithms and architectures, are the data used for training and testing the systems. Therefore, the availability of good datasets in which the researchers can train their algorithms and generate reproducible is of paramount importance. Aligned with the reproducible research trend, the authors of [7], [9] created a dataset useful for AMC. Also, in [10], new datasets are proposed with application of real channel impairments to generate realistic scenarios. As importance of machine learning increases, data plays a similar role to channel models, which are mandatory to the design of new RATs. In 5G development, for example, ongoing efforts are making data and channel models available [11].

The major contribution of this paper is a new publicly available dataset called *UFPAtelecom* for RATC that consists of LTE and GSM signals, both *digitized* and *artificially* generated (via simulation). We then used this RATC dataset and the AMC made available in [7], [9] to investigate the performance of classifiers comparing deep learning algorithms with traditional machine learning metrics in severely mismatched conditions, such as when they are trained and tested with data collected in indoor and outdoor environments to show the possible outcomes in those distinct scenarios. We also compare our results with others [7] and draw conclusions about possible sources of discrepancies and difficulties to perfectly reproduce experiments unless all data and software are properly organized.

This rest of this paper is organized as follows. In Section II, the *UFPAtelecom* dataset is described. In Section III, the feature extraction methods and classifiers used in this paper are briefly described. The results of simulations with the datasets are shown in Section IV. We conclude this paper in Section V.

II. UFPAtelecom Dataset

A. Motivation

Data are abundant or have a relative low cost in many machine learning application domains. For example, the text-to-speech system presented in [12], which represents the state-of-the-art, achieves quality close to natural human speech after being trained with 24.6 hours of digitized speech. In contrast, the research and development of machine learning for telecommunications has to deal with a relatively limited amount of data. The lack of freely available data impairs the data-driven lines of investigation. This is our primary motivation for the development of new datasets.

The second one is the importance that we believe needs to be placed on the propagation channels. Some of the currently available datasets include a large variety of signals with respect to the modulation, for example, but very limited variation of the channel. For example, in [7], a single channel model from the GNU Radio software was used to generate all signals, in both training and test subsets. However, in a communication system the channel is the only block that is not designed by humans. In contrast, the elaborate processing that happens in the brain to generate and detect speech is not under complete control of system designers as in telecommunications. Hence, the data "richness" depends primarily on the channel, given that input and output signals are learned along the machine learning process or determined by the adopted equipment and / or communication standard. For this reason, the philosophy behind the UFPAtelecom and other datasets being generated in our group prioritizes the variety of channels. We believe this approach will facilitate conclusions regarding the generalization capabilities of data-driven telecommunication algorithms.

With these two primary motivations, the *UFPAtelecom* dataset was created to help in telecommunications education, research and development of RATC algorithms. It can be downloaded from [13] and currently consists of LTE and GSM signals divided in three categories: **artificial** signals, signals digitized from commercial mobile networks owned by **operators** and software-defined radio (**SDR**) signals generated at our laboratory premises. Figure 1 describes the general organization of the dataset.

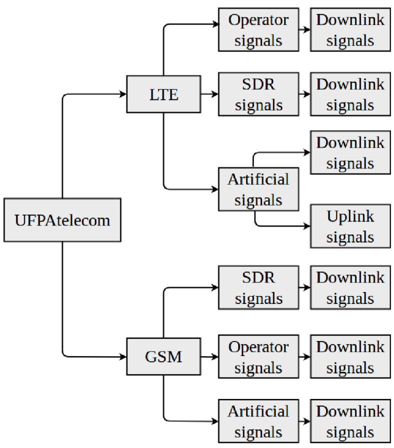

Fig. 1
ORGANIZATION OF LTE AND GSM MODULATION IN THE UFPATELECOM DATASET FOR ARTIFICIAL, DIGITIZED AND SDR SIGNALS.

The artificial signals indicated in Figure 1 were generated using the MATLAB software, by varying the signal-to-noise ratio (SNR) of a noise-free base signal. More specifically, we vary the SNR from -20 to 20 dB in steps of 2, creating 20 versions of a given base signal. The operator signals were collected using a Keysight EXA Signal Analyser N9010A (VSA) using the Keysight VSA 89600 software (VSA software). This hardware along with its software allows both offline and online demodulation and gives traces of the received signal that express data in time and frequency domains. The third category, the SDR signals, were created using a setup based on GNURadio with the Universal Software Radio Peripheral (USRP) SDR.

The information about the total size in Gigabytes (GB) and the total duration in seconds of the signals in the *UFPATelecom* dataset is shown in Table I. Details about each category of signals are presented in the next sections. Note that methods to properly obtain some uplink signals are still under development and there are no corresponding signals.

TABLE I
DURATION IN SECONDS (SEC) AND SIZE (GB) FOR UFPATELECOM SIGNAL CATEGORIES.

	LTE				GSM			
	Downlink		Uplink		Downlink		Uplink	
	Sec.	GB	Sec.	GB	Sec.	GB	Sec.	GB
Artificial	52.8	66	8.8	6.6	340	2.98	-	-
Operator	1.28	0.05	-	-	11.2	0.13	-	-
SDR	4800	199	800	20	17	0.33	-	-

B. Artificial (simulated) base signals

1) GSM: In order to create the GSM artificial base signals dataset, we utilized the Osmocom [14] software, which has an extensive library dedicated to GSM. With this software, a base GSM downlink signal is generated. This base signal is generated at 4 samples-per-symbol (sps), therefore the burst size is 625 instead of 156 bits, and has the duration of approximately 17 seconds (29,544 bursts), given that each GSM burst lasts for 0.577 ms. After the process of varying the SNR, the dataset has the total duration of 340 seconds and total size of 2.98 GB.

2) LTE: Both uplink and downlink artificial LTE base signals with bandwidths of 1.4, 3, 5, 10, 15, and 20 MHz were created using the MATLAB LTE toolbox [15]. After this process, the signal is passed through a multipath propagation channel. We selected the parameters defined by 3GPP [16] for the channel, and after that the Extended Pedestrian A model (EPA) [16] with Doppler frequency of 5 Hz was selected. Every base downlink signal of every bandwidth has 2.64 seconds of total duration, while the base uplink has 0.44 seconds of total duration. Taking in account every bandwidth utilized and the process of varying the SNR, the total duration of the dataset is 52.8 seconds and size of 66 GB for downlink and 8.8 seconds and size of 6.6 GB for uplink.

C. Mobile operators signals

As mentioned, the mobile operators signals were captured using VSA hardware and software. The measurements were performed at the building "Espaço Inovação" at UFPA. Both GSM and LTE digitized operators signals were searched within the frequency bands distributed among mobile operators by ANATEL, the National Agency of Telecommunication of Brazil, to operate in Amazon region.

1) GSM: The absolute radio-frequency channel number (ARFCN) channels were manually tuned with the help of the VSA hardware and software using the GSM/EDGE module. In the given scenario, only 7 downlink channels were found within the GSM 900 MHz frequency band. The module informs the GMSK constellation, burst type, demodulated slots, and also, calculates quality of service measurements regarding the received signal. These parameters were used to validate the signals, since we need to avoid recording only noise or other signals that are not GSM. The GSM digitized operators dataset has signals with frequency band of 200 KHz, sampling frequency of 640 KHz, and were recorded with frequency span value of 500 KHz which indicates the frequency bandwidth analyzed.

2) LTE: A similar effort was done to digitize operator's LTE signals, and only 5 downlink channels were received and analyzed. The LTE Advanced module of the VSA software was used to collect traces that show modulation constellation, demodulated physical channels values, active resource blocks, signal quality measures and others parameters. The LTE digitized operators dataset has signals with frequency band of 10 MHz.

D. Indoor software-defined radio (SDR) signals

Both GSM and LTE SDR signals, as mentioned in Section II, were generated using a GNURadio setup with two USRPs. The first USRP is used to transmit the noiseless artificial base signal to the air interface, while another is utilized to capture the signal sent by the first.

We used 560 LTE artificial signals, 480 for downlink and 80 for uplink, and 1 GSM artificial downlink signal.During the transmission and capture phase, the LTE signals were repeated until each signal reached 10 seconds of duration, while the GSM signal was only transmitted once. Hence, the total time for these signals is 4800 seconds for LTE downlink, 800 seconds for LTE uplink and 17 seconds for GSM downlink.

Both digitized GSM and LTE signals were stored as binary files. The real part of the first IQ sample is the first element of the file, and the imaginary part of the first IQ sample is the second element of the bin file and so on. We used a plain (raw) float numbers in little-endian format to help researchers to use them on different platforms.

III. Adopted Classifiers and Features

This section describes the adopted classifiers and features. All classifiers are trained and tested with N_{tr} and N_{te} examples, respectively. The input to the classifiers is a vector with K real-valued elements. This paper adopted two distinct sets of features to represent the signals. The first are simply the complex-valued samples of the time-domain signal as adopted in [7]. The second set are the *statistical* features adopted in [17], which we call here *knowledge-based* to emphasize they are not automatically learned but hand-designed. In [7], the time-domain features were compared with cyclic cumulants and the former led to better results. In this paper we used the more elaborate set of $K = 136$ features adopted in [17].

A. Features

1) Time-domain samples: In this paper, we compare the performance of AMC algorithms using versions of the RML2016.10a dataset presented in [7]. It should be noticed that the authors later released the RML2018.03 dataset, which is itself a modified version of RML2016.10a. For AMC we used these two datasets.

Given the well-known capability of CNNs to extract sensible features from raw data [7], the authors used in [7] a window of 128 consecutive complex-valued time-domain samples as input features for the deep neural networks. Hence, the classifier input feature vector has dimension $K = 256$ to account for the real and imaginary components. The signal generated for each modulation is passed through a simulated channel (a GNU Radio model) that includes random walk drifting of carrier frequency oscillator, multi-path fading of the channel impulse response and additive Gaussian white noise.

2) Knowledge-based features: As mentioned, in order to generate the classifier's input parameters, we adopted the extended set of features presented in [17]. For features extraction, the maximum squared magnitude frequency component $A[k]$ and the maximum value of a Discrete Fourier Transform ($X[k]$) are calculated in order to derive features that considers the ratio between two maxima in the absolute values of $X[k]$, the estimation of the center frequency according to the observation window size of samples, and the ratio between the signal powers within different bandwidths around an estimated center frequency. For lack of space, the reader is invited to refer to [18] for more details.

B. Classifiers

AMC and RATC consist in identifying the modulation scheme or the RAT of a given communication system with a high probability of success and in a short period of time. The exact values for this probability and sensing period depends on the application, especially if the processing can be done offline or not. Another important aspect of AMC and RATC is the *prior information* that a classifier has, such as the symbol rate, the signal bandwidth, carrier frequency, etc. In the literature of AMC and RATC, some algorithms are tested by assuming simplified channel models and/or perfectly estimated parameters, which may be unrealistic in several practical scenarios. In this paper we used not only distinct datasets, but also classifiers with different characteristics with respect to accuracy and computational cost.

We used four distinct machine learning algorithms in this paper. The first is the Decision Tree (DT), which can be seen as a set of if/else rules based on thresholds. To classify an instance to a class, the DT algorithm traverses the tree to find the leaf node for this instance and returns the ratio of training instances of the class in this node. The DT is considered a white box model due to its easy interpretability [19]. A related technique is the Random Forest (RF), which is an ensemble (more specifically using bagging) of several decision trees. The RF algorithm searches for the best feature among subset of features, instead of searching for the very best feature when splitting a node, which adds extra randomness when growing trees [19]. It yields great tree diversity and overall better model. Another traditional method for multiclass classification is the Naïve Bayes (NB) algorithm which applies the Bayes theorem considering the independence between every pair of feature, which characterizes a naïve assumption. The NB method can be very fast compared with other algorithms and helps to alleviate problems of dimensionality [20]. The last learning algorithm employed was the Convolutional Neural Network (CNN) which is composed of multiple hidden layers which differentiates from the traditional neural network by the convolutional layer that ensemble low-level feature from one layer to high-level features of consecutive layers reaching great performance [19].

IV. Results

We first present results for AMC using datasets *RML2016.10a* and *RML2018.03*. Recall that the $K = 256$ features adopted for experiments with these two datasets are the 128 complex-valued time-domain samples. We did not use the knowledge-based features in AMC. Then we discuss RATC using the *UFPATelecom* dataset with both time-domain and knowledge-based features.

A. AMC

The goal of our AMC experiments is to draw direct comparisons with [7]. We use the same convolutional neural network used in [7] but adopted three different dropout probabilities. The dropout helps the network generalization capability and is widely used in deep learning.

The following parameters were adopted for the three classifiers. For the DT, the maximum depth parameter was 80. For RF, the number of trees was 80 and the maximum depth was 200. For the CNN, the number of epochs was 100.

Table II shows the mean accuracy among all SNR values for each classifier tested in this work. It is possible to see that, in all cases except for RF, the tests with the *RML2018.03* dataset achieved better accuracy and most results are comparable to the ones obtained in [7], but not exactly the same. More importantly, the results in Table II lead to the conclusion that, while the CNNs can extract features from raw data, the time-domain samples are not reasonable features for classifiers such as decision trees, which are based on if/else rules and thresholds. This observation led us to design the following RATC experiments using also a set of improved knowledge-based features.

B. RATC

The total number N of examples available depend on the adopted features. For the time-domain samples, N can be easily calculated given the total number of samples and the

TABLE II

Overall AMC accuracy of classifiers for datasets *RML2016.10a* and *RML2018.03* using the time-domain samples as features.

Algorithms	**RML2016.10a**	**RML2018.03**
CNN2 Dropout (60%)	62.3%	63.6%
CNN2 Dropout (50%)	62.7%	67.0%
CNN2 Dropout (0%)	68.8%	69.1%
Decision tree	24.4%	25.4%
Naive Bayes	18.5%	20.2%
Random forest	13.6%	13.3%

TABLE III

Overall RATC accuracy using the knowledge-based as features.

Train dataset	Test dataset	CNN	DT	RF
50% *ar*	50% *ar*	90.00%	83.36%	89.82%
50% *ar*	50% *op*	47.69%	52.65%	41.706%
50% *ar*	50% *sdr*	52.39%	53.23%	45.124%
50% *ar* + 50% *op*	50% *ar*	88.56%	83.86%	88.98%
50% *ar* + 50% *op*	50% *op*	98.932%	99.976%	99.98%
50% *ar* + 50% *op*	50% *sdr*	96.32%	98.32%	97.9%

window duration. Similarly, for the knowledge-based features, each 512 samples of the raw data generates an example of 136 samples (each number of the samples represents a feature from Section III).

For RATC experiments, the values of N for each signal category when using the knowledge-based features are 560040, 25768 and 295090, for artificial (*ar*), operator (*op*) and SDR (*sdr*), respectively. Figure 2 shows the RATC accuracy result over SNR values when using time-domain samples as features, while Table III shows the RATC accuracy result when using knowledge-based features.

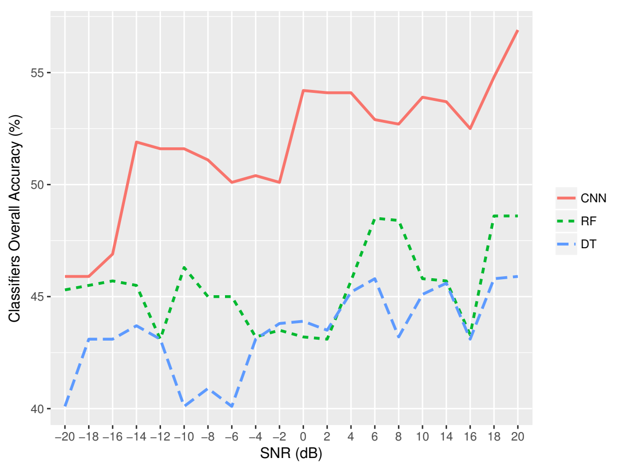

Fig. 2

RATC accuracy for test datasets with distinct SNR values using time-domain samples as features.

The RATC results using the time-sample as feature show that even CNN does not provide a good accuracy, in contrast to the AMC experiments here and in [7]. However, when the features from [17] were adopted, the accuracy increases for the artificial test set and, if the operator signals compose the train set by 50%, the accuracy for all the test sets returns a good accuracy.

Figure 3 provides detail about how the RATC classifiers perform for different SNR in the test sets. In this case the train dataset was composed only by artificial signals.

The combination of data from distinct sources in the training phase helps to evaluate the robustness of classifiers, in this case, our purpose is to illustrate how different can be the results in mismatched conditions. The Table III shows that RF and DT classifiers achieved similar performance in comparison with CNN, in the majority of scenarios that use indoor and outdoor data for training, which suggests that other factors besides performance should be evaluated in this case.

In fact, the DT algorithm is the approach with the lower computational cost and with fastest processing for the three scenarios, with a maximum depth $D = 80$, and also, RF with the number of trees $T = 80$ and depth $D = 200$ has a minor computational complexity in comparison with CNN, that has 3 dropout layers, 2 convolutional layers and 2 dense layers which demanded a major level of computation between layers and more processing time.

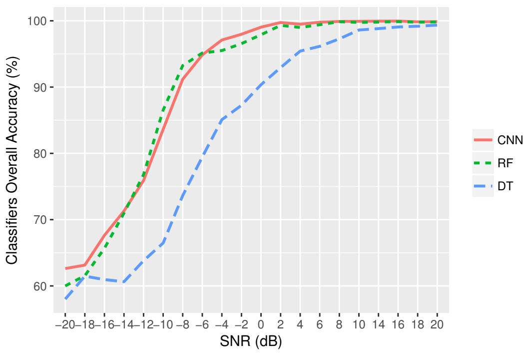

Fig. 3

RATC ACCURACY FOR TEST DATASETS WITH DISTINCT SNR VALUES AND THE KNOWLEDGE-BASED FEATURES.

V. CONCLUSIONS

This work presented the UFPATelecom dataset that can be used not only for RATC investigations, but also on distinct applications. Also, we have tested four classifiers applied to RATC using two distinct types of features. The first one was the time-sample features used in previous works, and the second was a knowledge-based set of features. This comparison allowed to conclude that the performance of many classifiers is highly-dependent on the appropriate choice of the features, and general conclusions require systematic evaluations using comprehensive combinations of features and classifiers. In addition, the classifiers outcomes revealed that traditional machine learning approaches can achieve good performance similar to CNN when trained with incompatible data. The results also emphasize how important is the channel in the assessment of machine learning applied to telecommunications. The datasets will be expanded taking in account this aspect.

REFERENCES

[1] O. A. Dobre, “Signal Identification for Emerging Intelligent Radios: Classical Problems and New Challenges,” *IEEE Instrumentation & Measurement Magazine*, vol. 18, no. 2, pp. 11–18, 2015.

[2] B. Tang, Y. Tu, Z. Zhang, and Y. Lin, “Digital Signal Modulation Classification With Data Augmentation Using Generative Adversarial Nets in Cognitive Radio Networks,” *IEEE Access*, vol. 6, pp. 15 713–15 722, 2018.

[3] S. Baban, D. Denkoviski, O. Holland, L. Gavrilovska, and H. Aghvami, “Radio Access Technology Classification for Cognitive Radio Networks,” in *2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2013.

[4] H. Cao, W. Jiang, M. Wiemeler, T. Kaiser, and J. Peissig, “A Robust Radio Access Technology Classification Scheme with Practical Considerations,” in *Personal, Indoor and Mobile Radio Communications (PIMRC Workshops), 2013 IEEE 24th International Symposium on*. IEEE, 2013, pp. 36–40.

[5] O. A. Dobre, R. Venkatesan, D. C. Popescu *et al.*, “Second-order Cyclostationarity of Mobile WiMAX and LTE OFDM signals and Application to Spectrum Awareness in Cognitive Radio Systems,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 1, pp. 26–42, 2012.

[6] Y. A. Eldemerdash, O. A. Dobre, O. Üreten, and T. Yensen, “Identification of Cellular Networks for Intelligent Radio Measurements,” *IEEE Transactions on Instrumentation and Measurement*, vol. 66, no. 8, pp. 2204–2211, 2017.

[7] T. J. O’Shea, J. Corgan, and T. C. Clancy, “Convolutional Radio Modulation Recognition Networks,” in *International Conference on Engineering Applications of Neural Networks*. Springer, 2016, pp. 213–226.

[8] X. Liu, D. Yang, and A. E. Gamal, “Deep Neural Network Architectures for Modulation Classification,” *arXiv preprint arXiv:1712.00443*, 2017.

[9] T. J. O’Shea and N. West, “Radio Machine Learning Dataset Generation with GNURadio,” in *Proceedings of the GNU Radio Conference*, vol. 1, no. 1, 2016.

[10] T. J. O’Shea, T. Roy, and T. C. Clancy, “Over-the-Air Deep Learning Based Radio Signal Classification,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 1, pp. 168–179, 2018.

[11] C. Beckham, “5G mmWave Channel Model Alliance,” 2016.

[12] J. S. et al, “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions,” in *https://arxiv.org/abs/1712.05884*, 2018.

[13] “UFPAtelecom Dataset.” [Online]. Available: https://www.lasse.ufpa.br/UFPAtelecom/

[14] “Open Source Mobile Communications.” [Online]. Available: https://osmocom.org/

[15] “LTE System Toolbox.” [Online]. Available: https://www.mathworks.com/products/lte-system.html

[16] “3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Base Station (BS) Radio Transmission and Reception (Release 8).” [Online]. Available: http://www.qtc.jp/3GPP/Specs/36104-820.pdf

[17] K. Lau, M. Salibian-Barrera, and L. Lampe, “Modulation Recognition in the 868 MHz Band Using Classification Trees and Random Forests,” *AEU-International Journal of Electronics and Communications*, vol. 70, no. 9, pp. 1321–1328, 2016.

[18] M. Kuba, K. Ronge, and R. Weigel, “Development and Implementation of a Feature-based Automatic Classification Algorithm for Communication Standards in the 868 MHz Band,” in *Global Communications Conference (GLOBECOM), 2012 IEEE*. IEEE, 2012, pp. 3104–3109.

[19] A. Géron, *Hands-on Machine Learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. "O’Reilly Media, Inc.", 2017.

[20] H. Zhang, “The Optimality of Naive Bayes,” *AA*, vol. 1, no. 2, p. 3, 2004.