NTL Detection: Overview of Classic and DNN-based Approaches on a Labeled Dataset of 311k Customers

Pablo Massaferro*, J. Matías Di Martino*†, and Alicia Fernández*
*Universidad de la República, Montevideo, Uruguay.
†Duke University, North Carolina, USA.
{pmassaferro, matiasdm, alicia}@fing.edu.uy

***Abstract*—Non-technical losses (NLT) constitute a significant problem for developing countries and electric companies. The machine learning community has offered numerous countermeasures to mitigate the problem. Yet, one of the main bottlenecks consists of collecting and accessing labeled data to evaluate and compare the validity of proposed solutions. In collaboration with the Uruguayan power generation and distribution company UTE, we collected data and inspected 311k costumers, creating one of the world's largest fully labeled datasets. In the present paper, we use this massive amount of information in two ways. First, we revisit previous work, compare, and validate earlier findings tested in much smaller and less diverse databases. Second, we compare and analyze novel deep neural network algorithms, which have been more recently adopted for preventing NLT. Our main discoveries are: (i) that above 80k training examples, the performance gain of adding more training data is marginal; (ii) if modern classifiers are adopted, handcrafting features from the consumption signal is unnecessary; (iii) complementary customer information as well as the geo-localization are relevant features, and complement the consumption signal; and (iv) adversarial attack ideas can be exploited to understand which are the main patterns that characterize fraudulent activities and typical consumption profiles.**

***Index Terms*—non-technical losses, electricity theft, automatic fraud detection,**

I. Introduction

Electric power is an essential asset for the society's development, and due to its distributed nature, has been vulnerable to theft and fraud [1], [2]. The verification of each customer power meter is untrackable, and therefore, smart ways of detecting potential fraud and non-technical losses (NTL) are essential. In the last decades, the machine learning community has offered several countermeasures to NTL [3], [4], [5], [6]. Yet, one of the main bottlenecks consists of collecting and accessing labeled data to evaluate and compare the validity of proposed solutions. In collaboration with the Uruguayan power generation and distribution company UTE, we collected data and inspected 311k costumers, creating one of the world's largest fully labeled datasets. The inspections were carried out by expert personnel over 6 years, and each client received at least one in place inspection by a certified electrician. The data was collected across the entire territory of Uruguay between January 2014 and July 2020.

In the present paper, we exploit the valuable collected data in two ways. First, we revisit previous work, compare, and validate earlier findings tested in much smaller and less diverse databases. Second, we compare and analyze novel deep neural network algorithms, which have been more recently adopted for preventing NLT [5], [6], [7]. We begin by describing the collected data and algorithms tested. Then we present and discuss the results obtained. We follow reviewing related work and conclude discussing our findings and perspectives for the future.

II. Methods

Data A novel fully label dataset of residential and commercial customers was collected in a joint effort between academia and industry. 311k clients were inspected by certified electricians to assess any sign of failure or fraudulent activities in the client electric installation and meter. The inspections were conducted through the entire country of Uruguay, between January 2014 and July 2020. For $36.6k$ out of the 311k inspections (11,8%), irregularities were detected. After inspection, if any abnormality is detected, we assign the label of *abnormal* to the customer, and otherwise, we label the customer as *normal*. In addition to the monthly power consumption and its label, we retrieve for each customer the additional information: *Contracted Power* represents the maximum power contracted by the client; *(latitude, longitude)* the geographical location of the meter; *Late Payment* the accumulated days of delay of bills payment; and *Fraud History* the number of previous irregularities detected among others [8]. A subset of the additional features is illustrated in Figure 2. To preserve customers anonymity, we truncated the latitude and longitude information to a 1km precision.

Feature extraction. Let us assume we study a set of n customers, for which we know: $C = (c_1, ..., c_m)$ the measured monthly consumption for m consecutive month. In the present work, we set $m = 36$ and defined the last consumption C_m as the one prior to the inspection. In addition to the monthly consumption,

complementary features are accessible, as described above, we denote these as $v = (v_1, ..., v_p)$. The ground truth label is represented as $y = 1$ or $y = 0$ for the positive (fraudulent) and negative class, respectively.

One of the earliest approaches to NTL consisted of handcrafting a set of features from the consumption curve [3], [9]. This step consists of defining a representation $(u_1, ..., u_k) = f(c_1, ..., c_m)$, mapping the input vector of monthly consumption into k features $\{u_i\}$. Examples evaluated in this work are (i) the consumption mean and standard deviation, (ii) the seasonal ratios, defined as the consumption ratio between last year season and the current one, (iii) Fourier coefficients (the first 5 are a common choice for a 3-year consumption signal), (iv) Wavelet Coefficients, (v) and the coefficients of a polynomial approximation.

Classification Considering as input the raw consumption data $(c_1, ..., c_m)$ or handcrafted features $(u_1, ..., u_k)$, the next step is to map the input X into a predicted label $\hat{y}$. (To define classification agnostic to the feature selection, denote from now on the input features as X.) Several classification techniques are popular for NTL detection. We evaluated the set of more frequent options: (i) Logistic Regression (LR) [10], [11], [12], (ii) Support Vector Machines (SVM) [10], [12], [13], [14], (iii) Random Forest (RF) [11], [15], (iv) Gradient Boosting (GB) [16], and (v) Extreme Gradient Boosting (XGB) [11], [12], [16].

DNN-Based classification. More recently, the remarkable success of deep learning transformed the field, and most strategies are shifting from the feature-extraction and classification paradigm to an end-to-end learning model. In this context, features and boundary decisions are jointly discovered in a data-driven fashion. In the present work, we tested the most popular and relevant DNN based alternatives: (i) a convolutional neural network (CNN), (ii) a recurrent long-short term memory network (LSTM), and (iii) a fully connected multi-layer network (MLP). Networks architecture is illustrated in Figure 1 and detailed in the Supplementary Material. We tested architectures trained exclusively on the consumption signal and networks trained to simultaneously exploit the consumption signal and the complementary features.

Adversarial attacks. One of the most challenging and exciting aspects of DNN solutions is understanding and interpreting what these models are learning. In contrast with handcrafted features, where we intuitively have patters in mind that we want to represent, neural networks learn optimal patterns in a data-driven fashion. To understand some of the features and patterns the models are learning, we implemented an adversarial attack algorithm as described in the following. Then, by synthetically *transforming* a normal consumption into a fraudulent one (and vice-versa), we can understand some of the patterns being captured by the models and contrast them with our intuition of the problem.

A linear perturbation method for adversarial attack is presented in [17]. Let $J(\theta, X, y)$ be the network cost function and θ the model parameters. Since the model is differentiable, we can adjust the input to produce a gradient descent/ascend in the model predicted output $\tilde{X} = X + \epsilon\, sign(\nabla_X J(\theta, X, y))$. ($\epsilon$) represent the perturbation step. An iterative application of this procedure can be exploited to transform the inputs until the model's prediction shifts the predicted category (see Results).

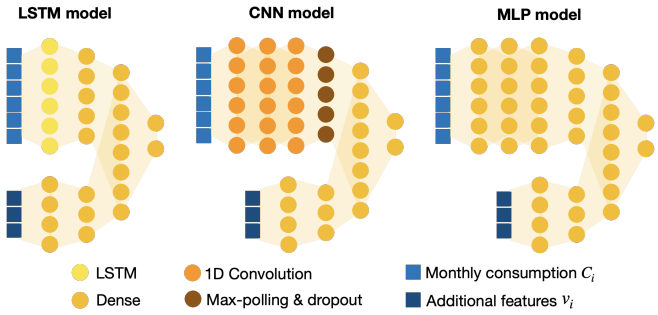

Fig. 1: DNN models tested. We tested three kinds of models, (i) a recurrent network with long-short term memory (left), (ii) a convolutional network (center), and (iii) a network with fully connected layers (right). For each model, we compared the performance learning exclusively from the consumption signal, and the combination of it with the additional features available.

Performance metrics. Being a problem with significant class imbalance, the evaluation of NTL models' performance is far from trivial. Because of this complexity, there is no consensus on the single more appropriate measure [1], [15]. In addition, the number of selected instances to be inspected is critical and depends on multiple external factors such as the inspection costs, and the operational capacity [8]. For this reason we report among other metrics, ranking measures such as the AUC_ROC (area under the true positive rate and falsepositive rate curve) and the area under the precision-recall curve AUC_PR. We report as well classical metrics such as the Recall, Precision, f-measure and Matthews Correlation Coefficient (MCC).

III. Results and Discussion

Additional features Figure 2 illustrates the distribution of the data across the country of Uruguay (a) and the distribution of a subset of additional features (b)-(e). The distance between the distributions associated with the positive and negative class is provided. As shown in Figure 2-(b), the payment delay distribution is shifted to the right for the positive class, suggesting that there is a positive correlation between a delay on the payment of the bill and the occurrence of fraud. Similarly, we observed that the history of fraud tends to be an indication of higher probabilities of fraud (Figure 2-(c)). In the case of the contracted power (Figure 2-(d)), a larger percentage of fraud is observed on the set of customer with lower contracted power. Another interesting observation is

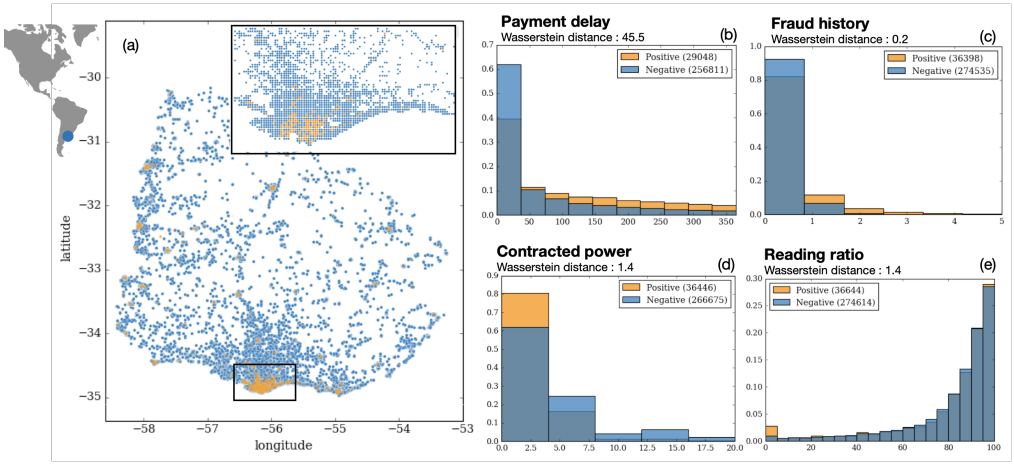

Fig. 2: Fully labeled data. (a) Geo-localization of a subset of the 311k labeled samples, in orange/blue, is illustrated in positive/negative samples. Plots (b)-(e) show the distribution of a set of the additional features across both classes; for each, the Wasserstein distance between the distributions of the positive and negative class is reported. The larger the distance, the larger the difference between the two distributions, meaning the more relevant the feature is to detect fraud. The number of samples used to estimate each distribution is provided for each plot . Reading ratio refers to the proportion of data obtained from the meter in site (in some cases where a meter reading cannot be accessed, the value is estimated by performing a regression with historical data).

the comparatively large number of fraudulent samples for which the Reading ratio feature is 0 (Figure 2-(e)).

Using Extreme Gradient Busting (XGB) as the classification model, we compared fraud detection performance when only the consumption curve is considered, and when we include additional features as described above (we compare classification algorithm in the following experiments). Figure 3 shows the precision-recall curves when XGB is trained exclusively with the consumption curve (dotted line), when we include the additional contract information such as the payment delay (dashed line), and finally, when we use all the information, including the geo-localization (solid line). This result validates similar results observed in the past [2], [15]. In this experiment, we only considered those customers for which 3 years of consumption information prior to the inspection data was available, from the 311k samples, 168k verified this condition.

Handcrafted features or raw data? Table I provides the results obtained for a set of classification strategies, trained to predict from the raw consumption curve or handcrafted features (we concatenated the features described in Methods). We can see that for all the algorithms used the results are superior when using raw data directly. We conclude then that extracting expert features is not the best way to work with NTL data. Since in previous experiments we observed no evident advantage of handcrafting features, in the rest of this section, we train and test our models using the raw consumption curve as input

Size of the training set. Since we had the unique opportunity of training with a very large dataset of over $150k$ examples, we evaluated how the size of the training set impacts performance. We observed that above $80k$ samples, the gain in performance becomes marginal (see Figure 3).

Algorithm	data	AUC_PR	AUC_ROC	Fmeasure	MCC
LR	Raw	25,5	64,7	32,0	0,094
	Features	19,5	61,9	29,6	0,073
SVM	Raw	25,4	64,5	31,9	0,092
	Features	19,4	61,7	29,6	0,074
RF	Raw	27,3	67,2	33,6	0,106
	Features	26,5	66,0	33,3	0,106
GB	Raw	27,2	67,5	33,9	0,117
	Features	27,1	67,1	33,7	0,116
XGB	Raw	26,9	66,5	33,5	0,110
	Features	26,5	65,7	32,5	0,102

TABLE I: Classification performance across several feature extraction and classification techniques. “Features” is a set of 30 concatenated hand crafted features detailed in Methods while “Raw” is the 36 normalized monthly consumption data.

What DNN-based models are learning? We train DNN models using the consumption curve and combining it with the additional features described in Methods. A schematic description of the models is provided in Figure 1. The three architectures evaluated (a recurrent, a convolutional, and a fully connected model) performed similarly. They outperform classical methods such as SVM, but present competitive results compared to modern alternatives such as XGB. In that sense, the question of whether

Algorithm	PR_AUC	ROC_AUC	Fm	MCC	Precision	Recall
MLP*	25,8	65,5	32,0	0,101	29,5	35,1
CNN*	**28,6**	**67,9**	34,1	0,116	33,0	35,3
LSTM*	28,2	66,3	34,2	0,241	35,5	32,3
LR	34,1	71,3	37,9	0,134	33,5	43,6
SVM	34,2	71,4	38,2	0,142	38,6	37,9
RF	38,8	75,2	41,3	0,155	36,1	48,3
GB	40,3	75,9	42,3	0,164	41,0	43,6
XGB	**40,5**	**76,2**	42,0	0,159	37,3	48,1
MLP	37,0	73,2	39,8	0,299	38,7	40,9
CNN	38,2	73,8	40,0	0,299	37,9	42,4
LSTM	38,4	73,6	40,4	0,301	37,3	44,2

TABLE II: Performance of all algorithms presented trained on consumption history and the additional data. The first 3 algorithms (*) were trained using only the consumption time series, complementing the results of Table I.

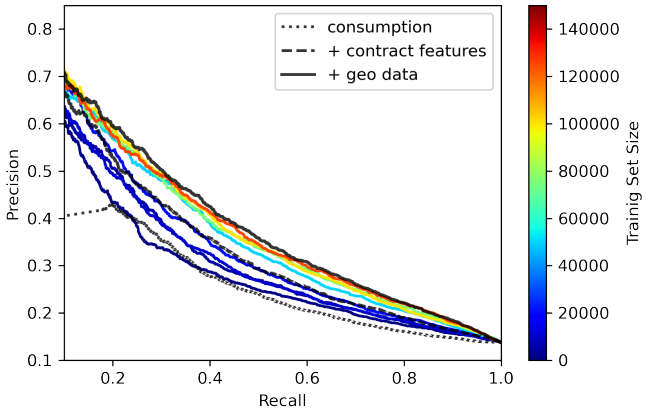

Fig. 3: Precision Recall Curve when additional contractual features and consumption data is combined, and when the size of the training set is modified. Training with different set of features is illustrated with different line styles, the dotted curve represent the performance of a model trained only on the consumption curve, the dashed line represent a model that includes additional information from the customer contract (see Methods), and finally, the solid line represent the accuracy when all the information (including the geo-localization) is considered. In addition, solid lines of different color represent the performance as we change the size of the training set. We started with 150k training samples (red) and decreased the size of the training set up to 10k samples (blue). This experiment was performed considering XGB as classification algorithm.

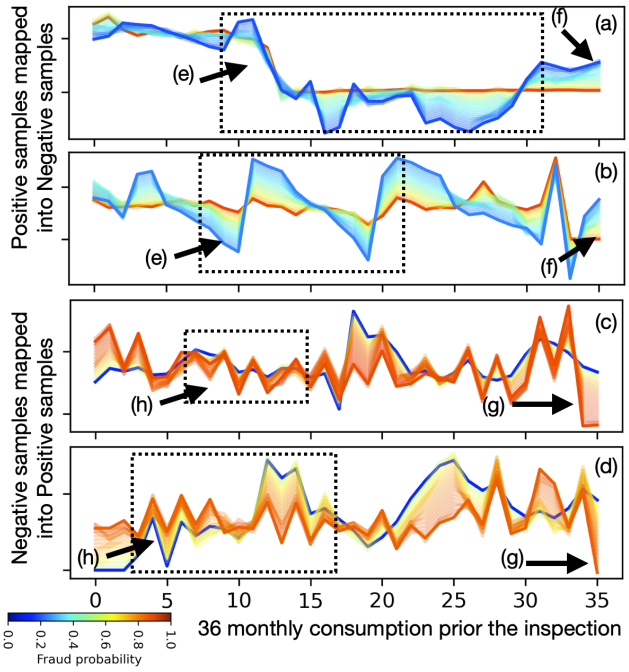

Fig. 4: Transforming a normal profile into an abnormal one and vice-versa. Examples (a-b) represent initially fraudulent curves (red curve in the background); as the adversarial attack proceeds (see Methods), we can observe how the curve evolves, and color becomes closer to blue (meaning the prediction shifts towards the normal class). Each curve is colored using the output of XGB model. In contrast, (c-d) show initially normal profiles (blue curves on the back), iteratively transformed into fraudulent profiles (red curves on the front).

DNN could potentially become the best tool in the future remains open. Testing more complex models (i.e., ones with more parameters and layers) and a close study of the optimization strategies constitute exciting future work.

As reported in previous work, performance improves when additional contractual geographical information is considered [2], [15]. To understand some of the patterns DNN-bases models are capturing, we used adversarial attack ideas to modify original data until the network prediction is modified (see Methods). Figure 4-(a-b) are examples of fraudulent samples transformed into typical ones. The opposite is illustrated in (c-d). Notice how transforming a profile into a fraudulent one typically involves a drop in consumption (Figure 4-(g)), and an atypical variability (Figure 4-(h)). On the other hand, to convert a fraudulent example into a typical one, a smooth seasonal variability is introduced (Figure 4-(e)), and rather than drooping, consumption oscillates (Figure 4-(f)).

IV. Related work

A revision of relevant works until 2018 is well detailed by Messinis et al. [1], where databases are categorized according to their size defined as large those with more than 1,000 records. Our work complements theirs by comparing a set of relevant solutions in a new, fully annotated database of over 300k customers in South America. In contrast with prior work, we do not assume that customers are normal by default [11], and only consider ground truth labels after a thoroughly in site inspection. We also compare novel DNN-bases approaches with classical ones, validating on a new and extensive database some previous findings [11], [13]. For example, we validated that categorical features are important and descriptive, showing significant Wasserstein distance between the distribution across classes. We also validated how geographical localization [15] is useful and complementary [10], [13]. There are very few works that present results on a dataset of our magnitude, and the present work complements their findings [10], [11], [12]. Current advances on DNN are providing remarkable power to extract features in a data-driven fashion, and our results suggest that previous handcrafted features [9] no longer provide clear advantages. Our unique findings are (i) that above $80k$ labeled customers, the performance increase seems to be marginal. This has huge practical implications when a company plans and designs data collection efforts; (ii) to compare on the same dataset classic approaches and DNN-based alternatives, and (iii) to use adversarial attack ideas as a way of learning what data-driven features extraction is capturing as meaningful indications of fraud and normality.

V. Conclusions

We presented and discussed a novel fully labeled dataset of over 311k customers collected across Uruguay. We compared classic approaches and handcrafted features with modern DNN-based methods. We observed that the latter could identify general patterns while optimizing decision boundaries in a unified and data-driven fashion. Similar performance was regarded across different network architectures (recurrent, convolutional, and fully connected). Future work includes comparing these architectures from a perspective of computational resources, comparing, for example, their performance as a function of the number of parameters and the training/prediction time. We exploited adversarial attack ideas to explore some of the most significant patterns that were capture by networks. Also, we observed that after approximately $80k$ labeled customers, the performance gain is marginal. This conclusion is likely connected to the fact that we are working with monthly consumption curves, which are low-dimensional compared to the data collected from smart meters. Since most of the infrastructure is switching to smart meters (actually in Uruguay 25% are smart meters), future work includes re-evaluating the strategies summarized in the present work in this new future scenario. The coexistence of measurement technologies is a challenge to explore new approaches with multi-resolution algorithms.

References

[1] G. M. Messinis and N. D. Hatziargyriou, "Review of non-technical loss detection methods," *Electric Power Systems Research*, vol. 158, pp. 250–266, 2018.

[2] P. Massaferro, H. Marichal, M. Di Martino, F. Santomauro, J. P. Kosut, and A. Fernández, "Improving electricity non technical losses detection including neighborhood information," in *2018 IEEE PES General Meeting (GM) - IEEE Power and Energy Society, Portland, Oregon, USA, 5-9 aug*. IEEE, 2018, pp. 1–5.

[3] R. Jiang, H. Tagaris, A. Lachsz, and M. Jeffrey, "Wavelet based feature extraction and multiple classifiers for electricity fraud detection," in *IEEE/PES Transmission and Distribution Conference and Exhibition*, vol. 3. IEEE, 2002, pp. 2251–2256.

[4] P. Glauner, J. Meira, P. Valtchev, R. State, and F. Bettinger, "The challenge of non-technical loss detection using artificial intelligence: A survey," *International Journal of Computational Intelligence Systems 10.1 (2017): 760-775.*, 2017.

[5] S. Li, Y. Han, X. Yao, S. Yingchen, J. Wang, and Q. Zhao, "Electricity theft detection in power grids with deep learning and random forests," *Journal of Electrical and Computer Engineering*, vol. 2019, 2019.

[6] T. Hu, Q. Guo, H. Sun, T.-E. Huang, and J. Lan, "Nontechnical losses detection through coordinated biwgan and svdd," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[7] Z. Zheng, Y. Yang, X. Niu, H.-N. Dai, and Y. Zhou, "Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 4, pp. 1606–1615, 2018.

[8] P. Massaferro, J. M. Di Martino, and A. Fernández, "Fraud detection in electric power distribution: An approach that maximizes the economic return," *IEEE Transactions on Power Systems*, vol. 35, no. 1, pp. 703–710, 2019.

[9] M. Di Martino, F. Decia, J. Molinelli, and A. Fernández, "Improving electric fraud detection using class imbalance strategies." in *International Conference on Pattern Recognition and Methods, 1st. ICPRAM.*, 2012, pp. 135–141.

[10] P. Glauner, A. Boechat, L. Dolberg, R. State, F. Bettinger, Y. Rangoni, and D. Duarte, "Large-scale detection of non-technical losses in imbalanced data sets," in *Innovative Smart Grid Technologies Conference (ISGT), 2016 IEEE Power & Energy Society*. IEEE, 2016, pp. 1–5.

[11] W. Hu, Y. Yang, J. Wang, X. Huang, and Z. Cheng, "Understanding electricity-theft behavior via multi-source data," in *Proceedings of The Web Conference 2020*, 2020, pp. 2264–2274.

[12] M. M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero, and A. Gómez-Expósito, "Detection of non-technical losses using smart meter data and supervised learning," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 2661–2670, 2018.

[13] P. Jokar, N. Arianpoo, and V. C. Leung, "Electricity theft detection in ami using customers' consumption patterns," *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 216–226, 2015.

[14] A. Jindal, A. Dua, K. Kaur, M. Singh, N. Kumar, and S. Mishra, "Decision tree and svm-based data analytics for theft detection in smart grid," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 3, pp. 1005–1016, 2016.

[15] P. Glauner, J. Meira, L. Dolberg, R. State, F. Bettinger, Y. Rangoni, and D. Duarte, "Neighborhood features help detecting electricity theft in big data sets," in *Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies.* IEEE, 2016.

[16] R. Punmiya and S. Choe, "Energy theft detection using gradient boosting theft detector with feature engineering-based preprocessing," *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 2326–2329, 2019.

[17] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.