2023 IEEE Texas Power and Energy Conference (TPEC) | 978-1-6654-9071-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/TPEC56611.2023.10078584

Energy Theft Detection Using the Wasserstein Distance on Residuals

Emran Altamimi*, Abdulaziz Al-Ali†, Qutaibah M. Malluhi*†, Abdulla K. Al-Ali*
*Department of Computer Science and Engineering
†KINDI Center for Computing Research
Qatar University, Doha, Qatar

***Abstract*—Detection of electricity theft improves the sustainability of the smart grid, helps electrical utilities mitigate their financial risks, and improves the overall management of resources. In this work, we utilize an LSTM neural network to forecast a given day's energy consumption and construct residuals. The residuals are then compared to previous residuals from normal days using the Wasserstein distance. If the Wasserstein distance for the residuals of a day exceeds a threshold, the day is highlighted to indicate suspected energy theft. Our framework can be built upon existing forecasting models with minimal computational overhead to calculate the Wasserstein distance. The framework is also highly explainable, which reduces the cost of false positives significantly. Our framework was evaluated using a public dataset and was able to detect six attack models of energy theft and faulty meters, with a false positive rate of 9% and an average F1 score of 0.91.**

***Index Terms*—Energy theft, Non-technical losses, Information distance, Wasserstein distance, Short-term load forecasting**

I. Introduction

Non-technical losses (NTL) are electricity losses that cannot be estimated or attributed to transmission and distribution. Non-billed electricity consumption, equipment breakdown, billing errors, low-quality infrastructure, and unlawful power consumption are all examples of non-technical losses [1]. Such losses incur a marginal drop in profitability for power utilities. NTL's adverse impact extends even beyond financial losses; for instance, it is considered the main culprit of uncertainty in countries' energy data, leading to the misallocation of resources and inefficient tariff design [2].

Smart grids offer an edge in the detection of NTL over traditional methods due to the higher resolution data that allows for advanced analytics [3]. One study approximates that smart grids can reduce the impact of NTL by 20% [4]. However, the authors in [5] surveyed the literature on NTL and found several challenges associated with its detection, such as the high cost of on-site inspection, which is generally not worth the utility provider's investment.

Recently, the work of [6] reviewed the attack models (either physical or cyber) and the supervised and unsupervised detection methods in the context of smart grids. The authors in [3] reviewed NTL detection systems that were grid-oriented and data-oriented. The grid-oriented methods make use of approaches such as state estimation to detect if the grid state is anomalous. While these methods are more accurate, they are typically harder to implement, more costly to maintain, and less reliable when an attacker manipulates the readings of several consumers in the same distribution area. Data-oriented approaches, on the other hand, are based on supervised and unsupervised learning algorithms. While supervised methods have demonstrated superior performance [3], they depend on the availability of labeled datasets, which are not common. To the best of our knowledge, there is no publicly available fine-grained labeled dataset for energy theft at this time of writing. Despite the existence of private labeled datasets [7]–[10], they however suffer from the class imbalance problem which is known to affect classification performance. Even worse, supervised methods struggle to classify new attack types that were not present in the dataset. Unsupervised methods in this case are preferable. However, their performance is suboptimal in comparison to supervised methods. Specifically, they lead to a higher false positive rate, which increases the cost of site inspection. False positives are instances where models are unable to differentiate between actual energy theft and genuine consumption, especially in instances where the consumer incurs genuine changes in their consumption patterns that are indiscernible for the models.

In this work, we propose a novel residual analysis method for detecting days on which energy theft occurred. NTL days are detected through two-sample hypothesis testing, which aims to determine if the two samples belong to the same or different underlying distributions. The Wasserstein distance is utilized because of its properties that are useful in the context of power consumption, which is discussed in section III.

II. Related Work

This section provides a comprehensive overview of related literature that uses regression and forecasting models to detect NTL, similar to our approach.

An Auto-Regressive Integrated Moving Average (ARIMA) model was utilized in [11] to forecast and mitigate electricity theft. The system detects electricity theft under a strong attack model where the adversary has full knowledge of the system and has full control over his own and his neighbor's smart meter readings. As a result, the attacker is overlooked during

This publication is supported in part by grant NPRP12C-33905-SP-66 and from the Qatar National Research Fund.
978-1-6654-9071-9/23/$31.00 ©2023 IEEE

the balance check, in which the utility provider ensures that the electricity supplied in a specific distribution network is correct, as discussed in [12]. The authors construct two sets that contain the means and standard deviations of weekly consumptions for each consumer. The mean and standard deviation of the following week's consumption must fall between the minimum and maximum values of the two sets.

The authors in [13] proposed a linear regression model to predict malicious consumers or faulty meters. The authors formulated the equation $\mathbf{y}_t = \mathbf{P}_t\mathbf{a}_n$ where $\mathbf{y}_t$ represents the difference between the supplied and reported consumption reading at the t_{th} time interval, $\mathbf{P}_t$ represents the reported consumption at the t_{th} time interval, and $\mathbf{a}_n$ represents the anomaly coefficient for the n_{th} consumer. To solve the equation for a_n the authors assumed $\mathbf{y}_t$ comes from a normal distribution of $\mathbf{N}(\mathbf{d}_{\mathbf{t}_\mathbf{i}}, \sigma^2)$ where $\mathbf{d}_{\mathbf{t}_\mathbf{i}}$ is $\sum_{n=1}^{N} a_n p_{t_i,n}$ such assumptions allow for maximum likelihood estimation to obtain a_n. The null hypothesis is thereafter tested using t-statistics, and the p-value is calculated with a threshold of 1% for each consumer. After obtaining the a_n and p-values, consumers with a higher p-value than the threshold are considered honest. Otherwise, a value greater than 1 a_n indicates possible electricity theft (i.e., reporting consumption), while less than 1 indicates a faulty meter (i.e., over-reporting consumption).

Our work is similar to the work of [11], [14], [15] where the residuals of the forecasting models are utilized to detect theft. However, we utilize the Wasserstein distance to classify electricity theft.

III. Proposed Framework

Our work analyzes the residuals of the forecasting errors to classify days as normal or as having NTL. An LSTM recurrent neural network is utilized to forecast the next sample reading for each consumer. The residuals in the forecasting step are then compared with the normal days' residuals using the Wasserstein distance. LSTM stands for Long-Short Term Memory, which is a type of recurrent neural network. The overall model consists of three layers: input, hidden, and output. The hidden layer consists of memory blocks; each memory block consists of three gates: an input gate, an output gate, and a forget gate. The model's architecture consists of a 128-neuron LSTM layer, a drop-out layer of 0.05, a 128-neuron dense layer, a second dropout layer of 0.05, a second dense layer, and a single output dense layer with a linear activation function.

The validation loss was monitored, and the model's parameters were saved at the lowest recorded validation loss.

The features selected and handcrafted for the LSTM model are the past day's readings, sub-hour (first, second, third, and fourth 15 minutes of each hour), hour, time-slice (each 4 hours segment of the day), weekday, day of the month, month and the difference from the previous value, the same reading at the same time on the day before (and a week before), the rolling average of the past day (and the past week). The model prediction of the next reading is then subtracted from the actual reading to form the residuals.

The Wasserstein distance measures the distance between two distributions, or "how different" they are from each other. The Wasserstein distance is bounded between $[0, \infty)$, such that the distance between two probability distributions is zero if and only if the two distributions are identical. The Wasserstein distance has some key advantages over other statistical distances when applied in the context of the smart grid [16], [17]:

- The Wasserstein distance can measure the distance between a continuous and a discrete probability distribution with a total variation of $1/N$, where N is the total number of samples, compared to 1 in other statistical measures. Such a property makes it possible to compare two probability distributions of various sizes with minimal variance.
- It can take into account the underlying geometry of space. while other statistical distance metrics ignore such a difference. which is of significance when dealing with electricity consumption. For example, two normal distributions with the same standard deviation but a different mean will result in a non-zero distance as opposed to some other metrics.
- The Wasserstein distance is not very sensitive to small differences in the distributions. Because residential consumption is noisy, it is critical to choose a metric that is not influenced by minor noise.
- Different weights can be assigned to empirical observations; certain observations, such as peak hours, can be assigned higher significance.

To understand how the Wasserstein distance is used in our approach, consider a given consumption pattern, Y_{real}. Here, an LSTM model is trained on the distribution of daily, weekly, and seasonal consumption patterns to forecast future consumption values $Y_{predictions}$. If we assume a perfect model and a day that is typical of the previous days, the model would produce predictions that are identical to the real observations. However, since there is no perfect model and the consumption pattern of consumers is prone to noise, we expect the errors to have a mean of zero with a certain, ideally small, variance. In other words, for a good model and given a day that exhibits the behaviors and patterns of the training days, the prediction errors will follow a normal distribution such that:

$$Y_{real} - Y_{predictions} = N(0, \sigma)$$

where σ is the standard deviation. If, however, the day exhibited patterns that were not accurately predicted by the model (i.e., atypical or anomalous), the residuals would be biased and have a different distribution. To quantify the difference between the two distributions, the residuals of each day are compared with the residuals of previous days (i.e., normal days) using the Wasserstein distance. Classification can then be done in several different ways. In our framework, we opted to compare normal days with each other to form a baseline. As such, the average of all the distances between the normal days and the standard deviation was calculated. We set the threshold

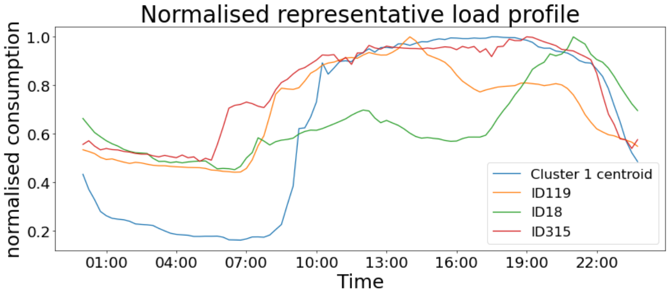

Fig. 1. Cluster 1 centroid and three consumers from the same cluster. Each line represents the normalized representative profile.

distance of anomalous days to be two standard deviations away from the baseline mean. This is based on the assumption that the Wasserstein distance on residuals between each pair of normal days follows a normal distribution, so any day that deviates from the typical residuals is considered anomalous. Then each new day is compared with all previous normal days, such that each normal day votes on whether the day of interest is anomalous or not. The final classification is then done through majority voting.

IV. Data-set and preprocessing

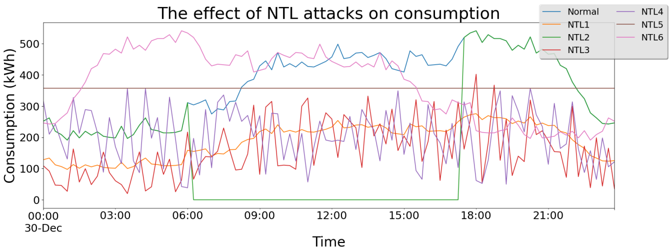

Fig. 2. The effect of NTL attacks on consumption.

A publicly available dataset, ElectricityLoadDiagrams201120145 [18], was used to validate the proposed methodology. It is made up of consumption data from 370 users between 2011 and 2014, with observations made every 15 minutes. The data were preprocessed using the following steps:

1) Consumers that had a significantly higher number of empty readings were removed. which resulted in a total of 315 consumers.
2) Since a significant proportion of consumers did not have data for the first year (2011), it was not considered in our evaluation.
3) Empty readings were then imputed utilizing KNN imputation.
4) Based on the elbow method, three clusters were identified, such that 20 consumers were randomly selected for evaluation, and the centroid was used to train the model that represents each cluster. The model is later used to make predictions about the 20 consumers.
5) The data was split into training and testing sets, where the testing set was the last 3 weeks of consumption for all consumers.
6) A MinMax scaler was fitted.

Figure 1 shows the normalized representative load profile of the cluster centroid and three randomly selected consumers for evaluation. The representative load profiles are the centroid of the daily consumption for each consumer. The three consumers have the lowest, median, and highest average consumption among the randomly selected 20 consumers. Note that the centroid is not a very good representative of all 20 randomly selected consumers. As such, the forecasting accuracy of a model trained on the centroid and tested on the cluster's randomly selected consumers is expected to be suboptimal. Admittedly, training an individualized model per consumer should provide better performance, albeit at the expense of a higher computational cost. However, we emphasize that this work targets the use of residuals to detect energy theft rather than constructing good short-term forecasting models. Having better forecasting models should increase our NTL detection approach.

To evaluate our model's performance on energy theft, the following procedure was done on each of the 20 consumers. The testing data was split chronologically into one day (to make the first prediction by the model), one week (for comparison with subsequent days), and 13 days for experimentation that were changed to evaluate our framework on energy theft. The consumption is then changed based on the six commonly used attack models, which were first proposed in [19] and then modified in [20]:

1) $t1(x_t) = x * random(0.1, 0.9)$
2) $t2(x_t) = x_t * random(0.1, 1.0)$
3) $t3(x_t) = x_t * random(0 or 1)$ for a random period
4) $t4(x_t) = mean(x) * random(0.1, 1.0)$
5) $t5(x_t) = mean(x)$
6) $t6(x_t) = x_{T-t}$ (where T is the sample size per day, that is, 96)

In addition, we consider an extra case to evaluate the model's capability to detect faulty meters. In this case, the readings were set to zero. Figure 2 shows the effects of the NTL attacks on consumption.

V. Results and Evaluation

The evaluation of the framework is split into two parts: the forecasting efficacy and distribution of observed residuals, and the capability of the Wasserstein distance to detect NTL days. Evaluating the forecasting performance of LSTM was done by calculating the RMSE. Recall that the LSTM model was trained only on the centroid of the whole cluster (which consisted of 314 consumers) to make predictions on 20 randomly selected consumers. The RMSE value on the testing set is 4.14 KWh, and the average consumption of the centroid is 352 KWh. Figure 3 shows the actual consumption, prediction, residuals, and probability distribution function (PDF) of the residuals. It can be noticed that the probability distribution of

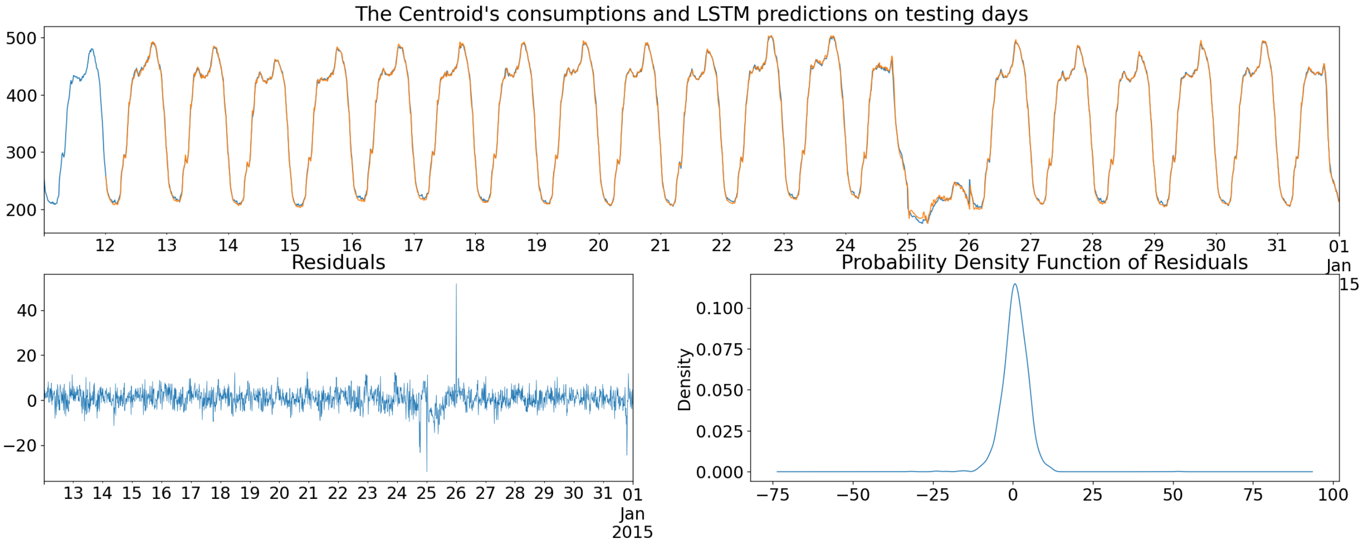

Fig. 3. The centroid's actual consumption, predictions, residuals (kWh against the 96 daily reported readings), and the PDF of residuals.

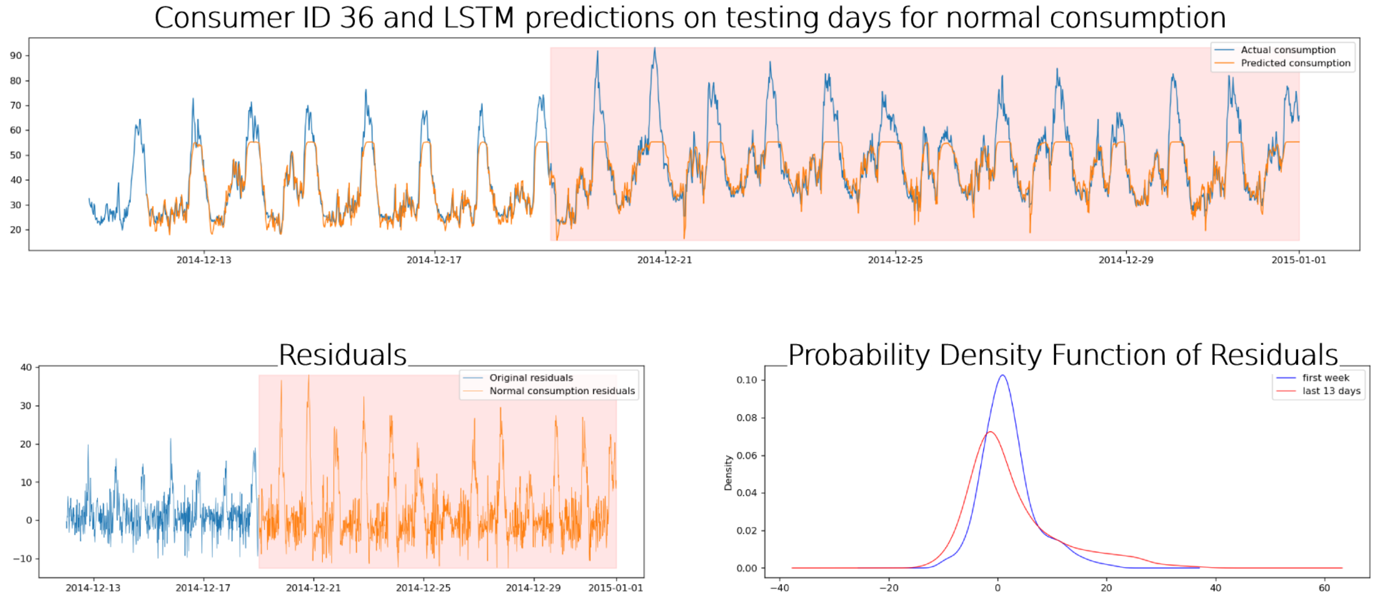

Fig. 4. Consumer ID 36 with the high false positives.

the residuals follows a Gaussian distribution with a mean close to zero, as expected.

Investigating the RMSE values of the random consumers along with the RMSE/average consumption ratio, which can be roughly interpreted as the average percentage error of the actual consumption, serves as a better indication of how well the model was able to predict consumption. All consumers (ID (18, 97, 119, 140, 167, 171,176, 197 211, 214, 224, 243, 260, 301, 315)) had an RMSE over mean consumption lower than 10 % except for four consumers. Three of which (ID (36, 66, 298)) had an RMSE over mean consumption of around 25 % and one (ID 300) of 46 %.

Next, we evaluate the false positive (false alarm) rate for the proposed method by classifying the original, unmodified consumption. Our framework had a false positive rate of 12% out of 400 normal days on unmodified data. However, upon further investigation, we found that consumer ID 36 had an abnormally high number of false positives: 13 false positives for that consumer, with the overall median being only 2 false positives. Figure 4 shows the consumer's consumption, the model's predictions, the residuals, and the PDF for both the first 7 days and the last 13 days. The model's failure to capture the peaks could be due to a systematic error in the LSTM model, and we argue that such a systematic error, if present, should be present across the entire dataset. As a result, the statistical distribution of residuals should be consis-

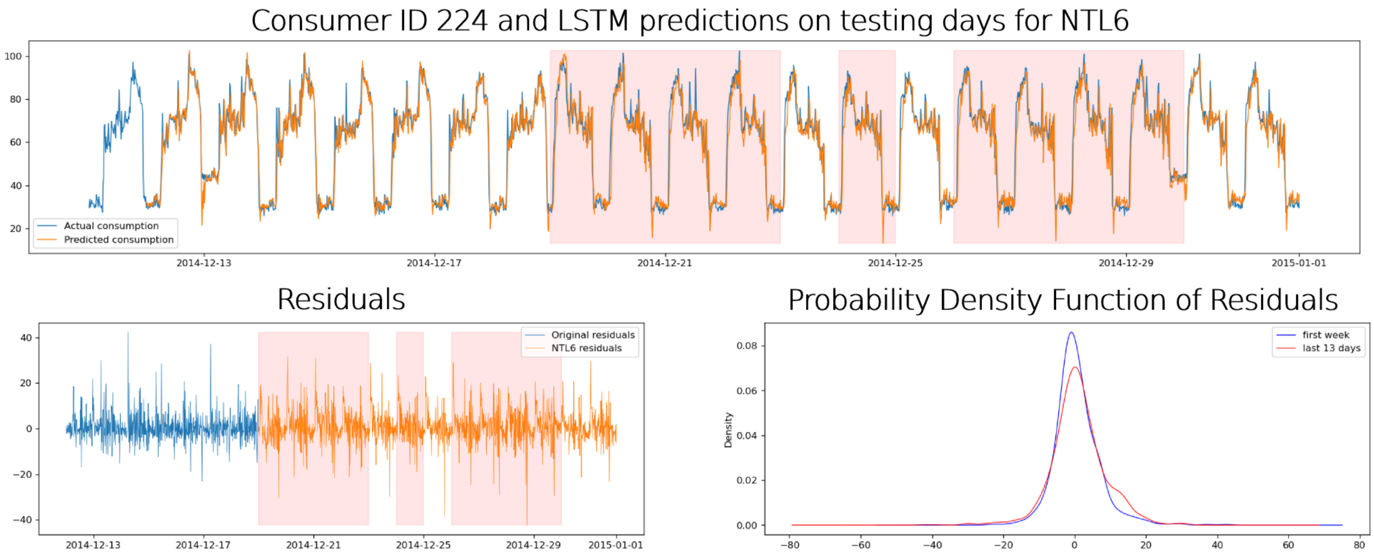

Fig. 5. Consumer ID 224 with NTL6.

TABLE I
THE EVALUATION SCORE OF THE MODEL ON NTL

Type	Accuracy	Precision	Recall	F1
NTL1	0.96	1.0	0.93	0.96
NTL2	0.95	1.0	0.93	0.96
NTL3	0.97	1.0	0.95	0.97
NTL4	0.98	1.0	0.97	0.98
NTL5	0.9	1.0	0.84	0.91
NTL6	0.6	1.0	0.38	0.56
Zeros	1.0	1.0	1.0	1.0

tent across the dataset.

We argue that the high positive rates for this consumer are due to an increase in consumption at the peaks. Since this is a real dataset, it is entirely possible that this consumer changed their consumption pattern. Removing that consumer reduces the average false positive rate to 9%.

The next experiments evaluated our model for NTL, or electricity theft. The results are summarized for every attack in Table I. While our framework performed almost perfectly in the first five attacks, it is apparent that it performed terribly in the 6th attack type (NTL6), which inverts the consumption values. This attack type reflects the case when the consumer reports lower consumption around peak hours to lower their electricity bill. Figure 5 illustrates consumer ID 224 with the NTL6 attack type. It can be noted that the residuals after NTL6 are almost as negative as the benign days.

To mitigate this, different weights can be applied to the residuals. While it is beyond the scope of this work to optimize or automate the selection of weights, simply changing the weights of the residuals for the first and last 6 hours of the day improved the performance to 0.68, 1.0, and 0.50 for accuracy, precision, and recall, respectively, with an F1 score of 0.68. During the investigation of the results for the poor performance of our framework on this attack type, we noticed a huge disparity in the results, where some consumers were classified with results that aligned with the previous results, and others were classified very poorly. for consumer ID 66, which was classified very poorly. We investigated the forecasting model's performance and found that it performed poorly at the consumption peaks. Accordingly, our approach failed to capture the discrepancies when the consumption was flipped, as the PDF of residuals was similar to what was deemed normal behavior.

VI. DISCUSSION

In this section, the advantages and limitations of our proposed framework are discussed while highlighting the tradeoffs that emerge from our framework. Our framework demonstrated exceptional results across six different attack types and the case of faulty meters. The framework achieved the results with a low false positive rate of 12 %. The framework is also easy to implement in real-world settings and is applicable to utility operators with minimal computational overhead to calculate the Wasserstein distance. The low complexity is attributed to the fact that the framework is a simple extension built upon already existing forecasting models. As demonstrated throughout the discussion of the results, our framework is highly explainable. This enables operators to decide whether a positive prediction justifies a site investigation or not, which can greatly reduce the cost of false positives. An unexplainable model or framework would result in each positive prediction having to be investigated in person. That being said, the utility can still, for example, require all the ”benign days” to vote that a day is anomalous for a day to be classified as anomalous, which would reduce false positives. On the other hand, it can require only one vote to indicate an anomaly, prioritizing recall. Another advantage is that our framework is robust against poor model performance. Despite the high RMSE for a couple of randomly selected consumers, the framework was still able to retain good results. We attribute this to the fact that, while the model performed badly on both the original and tampered days, it performed badly in ”different” ways, which was picked up by the Wasserstein distance.

A common limitation of previous works is that they rely on calculating the standard deviation of the residuals over sliding windows [11], [21]. The assumption is that a sudden change would induce a huge spike in the residuals, which would be larger than two standard deviations of the previous N residuals. Such frameworks will not be able to detect subtle changes in consumer behavior because typical ramping attacks realistically occur more subtly. In particular, each reading's residual will be within two standard deviations of the previous N readings. While such frameworks could use a fixed set of daily residuals for reference, the feasibility of such an idea has not been investigated. Our framework, however, is more robust against ramping attacks. While evaluating our framework against ramping attacks is beyond the scope of this work, we experimented on a random consumer where electricity consumption was incrementally scaled down by 5% per day to keep the underlying patterns of consumption intact so that the forecasting model performs just as well as on benign days. Our framework was able to achieve an F1 score of 0.76 when the scaling coefficient dropped to 0.70.

One limitation of our work is that the residuals are compared with the residuals of a set of benign days, which means that the set must be updated regularly and with the help of expert knowledge. Another limitation is that the framework cannot distinguish between NTL attacks and faulty meters, and it is up to the operator to decide.

VII. Conclusion

In this work, we propose a novel framework to detect electricity theft. The framework utilizes two-sample hypothesis testing to compare new days with predetermined days that are known to be normal. An LSTM model was trained on the centroid of the cluster and used to predict the consumption of randomly selected consumers within the cluster. The prediction errors are then compared between two days using the Wasserstein distance to determine whether a day is anomalous or not. Our framework demonstrated good results for six attack types and faulty meters. In future work, we aim to evaluate our framework in detecting concept drift, study the effect of changing the weights of the Wasserstein distance, and improve our framework's performance by training on the whole cluster or constructing individual models to reduce the RMSE values.

VIII. Acknowledgments

This publication is supported in part by grant NPRP12C-33905-SP-66 and from the Qatar National Research Fund. The findings achieved herein are solely the responsibility of the authors.

References

[1] T. Sharma, K. Pandey, D. Punia, and J. Rao, “Of pilferers and poachers: Combating electricity theft in india,” *Energy Research & Social Science*, vol. 11, pp. 40–52, 2016.

[2] Z. Sajid and A. Javaid, “A stochastic approach to energy policy and management: a case study of the pakistan energy crisis,” *Energies*, vol. 11, no. 9, p. 2424, 2018.

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

[4] H. Hamdan, R. Ghajar, and R. Chedid, “A simulation model for reliability-based appraisal of an energy policy: The case of lebanon,” *Energy Policy*, vol. 45, pp. 293–303, 2012.

[5] F. de Souza Savian, J. C. M. Siluk, T. B. Garlet, F. M. do Nascimento, J. R. Pinheiro, and Z. Vale, “Non-technical losses: A systematic contemporary article review,” *Renewable and Sustainable Energy Reviews*, vol. 147, p. 111205, 2021.

[6] M. G. Chuwa and F. Wang, “A review of non-technical loss attack models and detection methods in the smart grid,” *Electric Power Systems Research*, vol. 199, p. 107415, 2021.

[7] M. Di Martino, F. Decia, J. Molinelli, and A. Fernández, “Improving electric fraud detection using class imbalance strategies.” in *ICPRAM (2)*, 2012, pp. 135–141.

[8] J. Nagi, A. Mohammad, K. S. Yap, S. K. Tiong, and S. K. Ahmed, “Non-technical loss analysis for detection of electricity theft using support vector machines,” in *2008 IEEE 2nd International Power and Energy Conference*. IEEE, 2008, pp. 907–912.

[9] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and F. Nagi, “Improving svm-based nontechnical loss detection in power utility using the fuzzy inference system,” *IEEE Transactions on power delivery*, vol. 26, no. 2, pp. 1284–1285, 2011.

[10] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and A. Mohammad, “Detection of abnormalities and electricity theft using genetic support vector machines,” in *TENCON 2008-2008 IEEE region 10 conference*. IEEE, 2008, pp. 1–6.

[11] V. Badrinath Krishna, R. K. Iyer, and W. H. Sanders, “Arima-based modeling and validation of consumption readings in power grids,” in *International conference on critical information infrastructures security*. Springer, 2015, pp. 199–210.

[12] D. N. Nikovski, Z. Wang, A. Esenther, H. Sun, K. Sugiura, T. Muso, and K. Tsuru, “Smart meter data analysis for power theft detection,” in *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer, 2013, pp. 379–389.

[13] S.-C. Yip, K. Wong, W.-P. Hew, M.-T. Gan, R. C.-W. Phan, and S.-W. Tan, “Detection of energy theft and defective smart meters in smart grids using linear regression,” *International Journal of Electrical Power & Energy Systems*, vol. 91, pp. 230–240, 2017.

[14] V. B. Krishna, K. Lee, G. A. Weaver, R. K. Iyer, and W. H. Sanders, “F-deta: A framework for detecting electricity theft attacks in smart grids,” in *2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE, 2016, pp. 407–418.

[15] X. Wang, T. Zhao, H. Liu, and R. He, “Power consumption predicting and anomaly detection based on long short-term memory neural network,” in *2019 IEEE 4th international conference on cloud computing and big data analysis (ICCCBDA)*. IEEE, 2019, pp. 487–491.

[16] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde, “Optimal mass transport: Signal processing and machine-learning applications,” *IEEE signal processing magazine*, vol. 34, no. 4, pp. 43–59, 2017.

[17] C. Villani, *Topics in optimal transportation*. American Mathematical Soc., 2021, vol. 58.

[18] “Electricityloaddiagrams20112014.” [Online]. Available: https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014

[19] 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.

[20] 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.

[21] G. Fenza, M. Gallo, and V. Loia, “Drift-aware methodology for anomaly detection in smart grid,” *IEEE Access*, vol. 7, pp. 9645–9657, 2019.