People's Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

Ferhat Abbas University Setif 1



Faculty of Sciences

Department of Computer Science

Speciality: Cybersecurity

Master's Thesis

Evaluation of Deep Learning-Based Electricity Theft Detection Models Against Adversarial Evasion Attacks

Presented by:

Khadidja KEDDOUR

Nadia BOUDJEMLINE

Supervisor Dr. Sarra CHERBAL

President Dr. Linda KACHA

Examiner Dr. Bouchra MAATI

June 22, 2026

Academic Year: 2025–2026

Abstract

The increasing deployment of smart grids and smart meters has significantly improved the monitoring and management of electricity consumption. However, electricity theft remains a major challenge, causing substantial economic losses and affecting the reliability of power systems. Recently, machine learning and deep learning techniques have shown promising results in electricity theft detection. Nevertheless, these models are vulnerable to adversarial evasion attacks that can manipulate consumption data and degrade detection performance. This thesis evaluates the robustness of electricity theft detection models against adversarial attacks. First, electricity consumption data are preprocessed and used to train four baseline detection models: Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Multi-Layer Perceptron (MLP). The baseline performance of each model is evaluated using standard classification metrics. Subsequently, five adversarial evasion attacks, namely Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), Carlini and Wagner (C&W), Nearest Neighbor Perturbation (NNP), and Nearest Neighbor Distance (NND), are applied to generate adversarial examples and assess the vulnerability of the trained detectors. The performance degradation caused by each attack is analyzed and compared across all models. To explore potential improvements, a two-stage detector is proposed, where consumers are first clustered using K-means, and a fused ensemble of CNN, LSTM, GRU, and MLP models is then trained for each cluster and evaluated under the same attack scenarios. Experimental results show that while the proposed detector offers modest improvements against gradient-based attacks, particularly C&W, significant vulnerabilities remain, especially under dynamic perturbation attacks such as NND, which completely defeats both baseline and proposed models. This study highlights the importance of systematically evaluating robustness and the need for more fundamental advances to ensure secure smart grid operations.

Keywords: Smart Grid, Electricity Theft Detection, Deep Learning, Evasion Attacks, Cluster-Based Ensemble Learning.

Résumé

Le déploiement croissant des réseaux électriques intelligents (smart grids) et des compteurs intelligents (smart meters) a considérablement amélioré la surveillance et la gestion de la consommation d'électricité. Cependant, le vol d'électricité demeure un défi majeur, entraînant d'importantes pertes économiques et affectant la fiabilité des systèmes électriques. Récemment, les techniques d'apprentissage automatique (machine learning) et d'apprentissage profond (deep learning) ont montré des résultats prometteurs dans la détection du vol d'électricité. Néanmoins, ces modèles sont vulnérables aux attaques d'évasion adversariales qui peuvent manipuler les données de consommation et dégrader les performances de détection. Cette thèse évalue la robustesse des modèles de détection du vol d'électricité basés sur l'apprentissage profond face aux attaques d'évasion adversariales. Tout d'abord, les données de consommation électrique sont prétraitées et utilisées pour entraîner quatre modèles de détection de base : le réseau de neurones convolutionnel (CNN), la mémoire à long court terme (LSTM), l'unité récurrente à portes (GRU) et le perceptron multicouche (MLP). Les performances de base de chaque modèle sont évaluées à l'aide de métriques de classification standard. Par la suite, cinq attaques d'évasion adversariales, à savoir Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), Carlini and Wagner (CW), Nearest Neighbor Perturbation (NNP) et Nearest Neighbor Distance (NND), sont appliquées afin de générer des exemples adversariaux et d'évaluer la vulnérabilité des détecteurs entraînés. La dégradation des performances causée par chaque attaque est analysée et comparée pour l'ensemble des modèles. Afin d'explorer des améliorations potentielles, un détecteur en deux étapes est proposé, dans lequel les consommateurs sont d'abord regroupés à l'aide de l'algorithme K-means, puis un ensemble fusionné composé de modèles CNN, LSTM, GRU et MLP est entraîné pour chaque cluster et évalué dans les mêmes scénarios d'attaque. Les résultats expérimentaux montrent que, bien que le détecteur proposé apporte des améliorations modestes face aux attaques basées sur le gradient, en particulier CW, des vulnérabilités significatives subsistent, notamment face aux attaques à perturbation dynamique telles que NND, qui neutralisent complètement aussi bien les modèles de base que le modèle proposé. Cette étude souligne l'importance d'une évaluation systématique de la robustesse ainsi que la nécessité d'avancées plus fondamentales pour garantir un fonctionnement sécurisé des réseaux électriques intelligents.

Mots-clés: Réseau Électrique Intelligent, Détection du Vol d'Électricité, Apprentissage Profond, Attaques d'Évasion, Apprentissage par Ensemble Basé sur le Clustering.

ملخص

أدى الانتشار المتزايد للشبكات الكهربائية الذكية والعدادات الذكية إلى تحسين مراقبة وإدارة استهلاك الكهرباء بشكل ملحوظ. ومع ذلك، لا يزال سرقة الكهرباء يمثل تحدياً رئيسياً، لما يسببه من خسائر اقتصادية كبيرة وتأثيرات سلبية على موثوقية أنظمة الطاقة. وفي السنوات الأخيرة، أظهرت تقنيات التعلم الآلي والتعلم العميق نتائج واعدة في مجال كشف سرقة الكهرباء. ومع ذلك، تبقى هذه النماذج عرضة لهجمات التهرّب العدائية التي يمكنها التلاعب ببيانات الاستهلاك وتقليل فعالية أنظمة الكشف. تهدف هذه الأطروحة إلى تقييم متانة نماذج كشف سرقة الكهرباء المعتمدة على التعلم العميق في مواجهة هجمات التهرّب العدائية. في البداية، تتم معالجة بيانات استهلاك الكهرباء مسبقًا واستخدامها لتدريب أربعة نماذج أساسية للكشف، وهي: الشبكة العصبية الالتفافية ،(CNN) وشبكة الذاكرة طويلة وقصيرة المدى ،(LSTM) ووحدة التكرار ذات البوابات ،(GRU) والإدراك متعدد الطبقات .(MLP) ويتم تقييم الأداء الأساسي لكل نموذج باستخدام مقاييس التصنيف القياسية. بعد ذلك، تُطبَّق خمس هجمات تهرّب عدائية، وهي: طريقة إشارة التدرج السريع ،(FGSM) والطريقة التكرارية الأساسية ،(BIM) وهجوم كارليني واغنر ،(CW) وتشويش أقرب جار ،(NNP) ومسافة أقرب جار ،(NND) وذلك لتوليد أمثلة عدائية وتقييم مدى هشاشة النماذج المدرَّبة. كما يتم تحليل ومقارنة تدهور الأداء الناتج عن كل هجوم عبر جميع النماذج. ولاستكشاف تحسينات محتملة، يُقترح كاشف مكوَّن من مرحلتين، حيث يُجَمَّع المستهلكون أولاً باستخدام خوارزمية ،K-means ثم يُدرَّب لكل عنقود تجميعٌ مدمج من نماذج CNN وLSTM وGRU وMLP، ويُقيَّم ضمن سيناريوهات الهجوم نفسها. وتُظهر النتائج التجريبية أنه على الرغم من أن الكاشف المقترح يحقق تحسينات متواضعة في مواجهة الهجمات المعتمدة على التدرج، ولا سيما هجوم ،CW فإن ثغرات مهمة لا تزال قائمة، خاصةً في ظل هجمات الاضطراب الديناميكي مثل ،NND التي تقضي تمامًا على فعالية كلٍّ من النماذج الأساسية والنموذج المقترح. وتؤكد هذه الدراسة أهمية التقييم المنهجي للمتانة والحاجة إلى تطوير أساليب أكثر جوهرية لضمان التشغيل الآمن للشبكات الكهربائية الذكية.

الكلمات المفتاحية: الشبكة الكهربائية الذكية، كشف سرقة الكهرباء، التعلّم العميق، هجمات التهرّب، التعلّم التجميعي المعتمد على العناقيد.

Contents

Abstract **4**

Résumé **5**

6 **ملخص**

List of Figures **10**

List of Tables **11**

List of Acronyms **12**

General Introduction **14**

1 State of the art **16**

1.1 Introduction 16

1.2 Smart Grids and Electricity Theft 16

1.2.1 Smart Grid 16

1.2.2 Role of Smart Meters 17

1.2.3 Electricity Theft: Definition and Types 18

1.3 Machine Learning and Deep Learning for Electricity Theft Detection 19

1.3.1 Machine Learning Approaches 19

1.3.1.1 Supervised Methods 19

1.3.1.2 Unsupervised Methods 20

1.3.2 Deep Learning Approach 20

1.4 Adversarial Machine Learning 21

1.4.1 Definition of adversarial attacks 21

1.4.2 Evasion attacks 22

1.5 Defense Mechanisms Against Evasion Attacks 23

1.6 Conclusion 24

2 Related Work and Comparative Study **25**

2.1 Introduction 25

2.2 Related work 25

2.2.1 Summary of Limitations . . . 31
2.3 Conclusion . . . 31

3 Evaluation Framework and Proposed Methodology 32
3.1 Introduction . . . 32
3.2 Experimental Framework for Robustness Assessment . . . 32
3.2.1 Dataset Preprocessing . . . 32
3.2.1.1 Dataset Description . . . 32
3.2.1.2 Customer Selection and Train-Test Partitioning . . . 33
3.2.1.3 Data Cleaning and Validation . . . 33
3.2.1.4 Time-Series Sequence Construction . . . 33
3.2.1.5 Electricity Theft Simulation . . . 33
3.2.1.6 Data Standardization . . . 34
3.2.1.7 Dataset Labeling and Balancing . . . 34
3.2.1.8 Final Dataset Generation . . . 35
3.2.1.9 Dataset Summary . . . 35
3.2.2 Baseline Model Training . . . 36
3.2.2.1 Model Architectures and Hyperparameters . . . 37
3.2.2.2 Training Curves . . . 38
3.2.2.3 Threat Model . . . 39
3.2.3 Adversarial Attack Generation . . . 40
3.2.3.1 Mathematical Formulation of Evasion Attacks . . . 40
3.2.3.2 FGSM – Fast Gradient Sign Method . . . 40
3.2.3.3 BIM – Basic Iterative Method . . . 41
3.2.3.4 C&W – Carlini & Wagner Attack . . . 41
3.2.3.5 NNP – Nearest Neighbor Perturbation . . . 42
3.2.3.6 NND – Nearest Neighbor Distance . . . 43
3.2.4 Adversarial Sample Injection Ratio . . . 43
3.2.4.1 Fixed Attack Parameters . . . 44
3.2.5 Performance Evaluation Metrics . . . 44
3.2.6 Results . . . 45
3.2.6.1 Baseline Performance (Clean Test Set) . . . 45
3.2.6.2 Post-Attack Performance . . . 46
3.2.6.3 Vulnerability Analysis per Model . . . 48
3.2.6.4 Summary of Vulnerability Findings . . . 53
3.3 The Proposed Detector . . . 54
3.3.1 Motivation for the Proposed Architecture . . . 54
3.3.2 Design and Technical Architecture . . . 54
3.3.3 Cluster Selection and Customer Distribution . . . 57
3.3.4 Resilience and Performance Trade-off . . . 57

3.3.5 Discussion 64
3.3.5.1 Interpretation of the Results 64
3.3.5.2 Comparison with Related Work 64
3.3.5.3 Limitations 65
3.4 Conclusion 66

General Conclusion **67**

Bibliography **69**