TD 06/2024

FEDERAL UNIVERSITY OF PARÁ
INSTITUTE OF TECHNOLOGY
POSTGRADUATE PROGRAM IN ELECTRICAL ENGINEERING

LUCAS DE LIMA BASTOS

CLASSIFICATION AND CHARACTERIZATION METHODS OF NON-TECHNICAL LOSSES ON SMART GRID SCENARIOS

UFPA / ITEC / PPGEE
Guamá University Campus
Belém-Pará-Brazil

2024

TD 06/2024

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

"CLASSIFICATION AND CHARACTERIZATION METHODS OF NON-TCHNICAL LOSSES ON SMART GRID SCENARIOS"

AUTOR: **LUCAS DE LIMA BASTOS**

TESE DE DOUTORADO SUBMETIDA À BANCA EXAMINADORA APROVADA PELO COLEGIADO DO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA, SENDO JULGADA ADEQUADA PARA A OBTENÇÃO DO GRAU DE DOUTOR EM ENGENHARIA ELÉTRICA NA ÁREA DE COMPUTAÇÃO APLICADA.

APROVADA EM: 28/03/2024

BANCA EXAMINADORA:

Documento assinado digitalmente
gov.br **Eduardo Coelho Cerqueira**
Data: 02/04/2024 08:31:46-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Eduardo Coelho Cerqueira
(Orientador – PPGEE/UFPA)

Documento assinado digitalmente
gov.br **DENIS LIMA DO ROSARIO**
Data: 01/04/2024 14:35:00-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Denis Lima do Rosário
(Coorientador - PPGEE/UFPA)

Documento assinado digitalmente
gov.br **MARCOS CESAR DA ROCHA SERUFFO**
Data: 01/04/2024 14:21:50-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Marcos César da Rocha Seruffo
(Avaliador Interno - PPGEE/UFPA)

Documento assinado digitalmente
gov.br **IAGO LINS DE MEDEIROS**
Data: 01/04/2024 13:14:02-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Iago Lins de Medeiros
(Avaliador Externo ao Programa - PESQUISADOR/UFPA)

Documento assinado digitalmente
gov.br **ALLAN DOUGLAS BENTO DA COSTA**
Data: 01/04/2024 13:42:37-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Allan Douglas Bento da Costa
(Avaliador Externo - UFRA)

Documento assinado digitalmente
gov.br **THAIS LIRA TAVARES DOS SANTOS**
Data: 01/04/2024 11:54:25-0300
Verifique em https://validar.iti.gov.br

Prof.ª Dr.ª Thais Lira Tavares dos Santos
(Avaliadora Externa - PROCURADORIA GERAL DO PARÁ)

VISTO:

Documento assinado digitalmente
gov.br **Diego Lisboa Cardoso**
Data: 02/04/2024 10:12:28-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Diego Lisboa Cardoso
(Coordenador do PPGEE/ITEC/UFPA)

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)

D278c de Lima Bastos, Lucas.
CLASSIFICATION AND CHARACTERIZATION METHODS OF NON-TCHNICAL LOSSES ON SMART GRID SCENARIOS / Lucas de Lima Bastos. — 2024.
74 f. : il. color.

Orientador(a): Prof. Dr. Eduardo Coelho Cerqueira
Coorientador(a): Prof. Dr. Denis Lima do Rosário
Tese (Doutorado) - Universidade Federal do Pará, Instituto de Tecnologia, Programa de Pós-Graduação em Engenharia Elétrica, Belém, 2024.

1. Smart Meters. 2. Smart Grid. 3. Information Theory Quantifiers. 4. Non-technical Losses. 5. Ensemble Learning. I. Título.

CDD 607

TD 06/2024

FEDERAL UNIVERSITY OF PARÁ
INSTITUTE OF TECHNOLOGY
POSTGRADUATE PROGRAM IN ELECTRICAL ENGINEERING

LUCAS DE LIMA BASTOS

CLASSIFICATION AND CHARACTERIZATION METHODS OF NON-TECHNICAL LOSSES ON SMART GRID SCENARIOS

Thesis proposal submitted to the PhD. in Electrical Engineering at Federal University of Pará.

Advisor: Dr. Eduardo Coelho Cerqueira

UFPA / ITEC / PPGEE
Guamá University Campus
Belém-Pará-Brazil

2024

LUCAS DE LIMA BASTOS

CLASSIFICATION AND CHARACTERIZATION METHODS OF NON-TECHNICAL LOSSES ON SMART GRID SCENARIOS

Thesis proposal submitted to the PhD. in Electrical Engineering at Federal University of Pará.

Approved in: __/__/____

JUDGING PANEL

Ph.D. Eduardo Coelho Cerqueira
(Federal University of Pará - Advisor)

Prof. Dr. Denis Lima do Rosario
(Federal University of Pará - Internal Member)

Prof. Dr. Marcos Cesar da Rocha Seruffo
(Federal University of Pará - Internal Member)

Prof. Dr. Iago Lins de Medeiros
(Federal University of Pará - Internal Member)

Prof. Dr. Allan Douglas Bento da Costa
(Federal Rural University of Amazon - External Member)

Prof. Dra. Thaís Lira Tavares dos Santos
(Procuradoria-Geral do Estado do Pará - External Member)

Abstract

Classification and Characterization Methods of Non-Technical Losses on Smart Grid Scenarios

Advisor: Eduardo Coelho Cerqueira
Keywords: Smart Meters; Smart Grid; Non-technical Losses; Ensemble Learning; Information Theory Quantifiers.

Nowadays, grid resilience as a feature has become non-negotiable, significantly when power interruptions can impact the economy and society. Smart Grids (SGs) widespread popularity enables an immense amount of fine-grained electricity consumption data to be collected. However, risks can still exist in the Smart Grid (SG), since SG systems exchange valuable data, the distribution system loses substantial electrical energy. We divide this loss into two categories: technical and non-technical loss. A substantial amount of electrical energy is lost throughout the distribution system, and these losses are divided into two types: technical and non-technical. Non-technical losses (NTL) are any electrical energy consumed that is not invoiced. They may occur due to illegal connections, fraudulent activities, issues with energy meters such as delay in the installation or reading errors, contaminated, defective, or non-adapted measuring equipment, very low valid consumption estimates, faulty connections, and disregarded customers. Non-technical losses are the primary cause of revenue loss in the SG. Annually, electrical utilities incur billions in losses due to non-technical reasons. This thesis presents two detection methods of NTL: classification and characterization. We create an ensemble predictor-based time series classifier to classify NTL detection. This predictor uses the user's energy consumption as a data input for classification, from splitting the data to executing the classifier. Also, it assumes the temporal aspects of energy consumption data during the pre-processing, training, testing, and validation stages. The classification method has the advantage of classifying heterogeneous features in data. The characterization method proposes a study based on Information Theory Quantifiers (ITQ) to mitigate this challenge. First, we use a sliding window to convert the user's energy consumption time series into a Bandt-Pompe (BP) probability distribution function. Then, we extract the used ITQ. Finally, we apply each metric to the Probability Density Function (PDF) and map the layers to characterize their behavior. The characterization method is advantageous to be used when we have big data. Overall, our best results have been recorded in the fraud detection-based time series classifiers (TSC) model, improving the empirical performance metrics by 10% or more over the other developed models. Our results show that users with normal and abnormal energy consumption can be distinguished using only Information Theory Quantifiers by considering the range of values for each metric.

Contents

List of Abbreviations

ADB	ADaptive Boosting
BP	Bandt-Pompe
CAT	CATegorical boosting
CatBoost	Categorical Boosting
catch22	22 Canonical Time-series Characteristics
CECP	Complexity-Entropy Causuality Plane
DR	detection rate
DT	Decision Tree
ET	Extra Trees
FI	Fisher Information
FPR	false positive rate
GBCs	Gradient Boosting Classifiers
GBTD	Gradient Boosting Theft Detector
IEDs	Intelligent Electronic Devices
ITQ	Information Theory Quantifiers
k-NN	k-Nearest Neighbour
LGB	Light Gradient Boosting
LightGBM	Light Gradient Boosting Method

MCDCNN	Multi-Channel Deep Convolutional Neural Network
ML	Machine Learning
NTL	Non-technical Losses
OPF	optimum-path forest
PE	Permutation Entropy
ResNet	Residual Network
RF	Random Forest
SC	Statistical Complexity
SG	Smart Grid
SM	Smart Meter
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
t-LeNet	Time Le-Net
TL	Technical Losses
TS	Time Series
TSC	Time Series Classifier
TSF	Time Series Forest
XGB	extreme gradient Boosting
XGBoost	Extreme Gradient Boosting
XGBoost	Extreme Gradient

List of Figures

List of Tables

CHAPTER 1

Introduction

This thesis proposal presents classification and characterization methods of Non-technical Losses (NTL) in Smart Grid scenarios using Machine Learning (ML) and Information Theory Quantifiers (ITQ) . This chapter introduces the main ideas regarding the application of NTL and its challenges in Smart Grid (SG) scenarios, motivates this research work, and establishes the research questions, objectives, contributions, and text organization.

1.1 Overview

The advent of Smart Meters (SMs) technology heralds a new era in energy management, with its primary benefits being the enhancement of grid resilience and environmental performance [1]. In today's energy landscape, the ability of SG to ensure flexibility and reliability is paramount, especially given the economic implications of power interruptions [2]. SG accomplishes this by integrating additional dispersed and distributed power sources, facilitating the smooth incorporation of new resources, and providing corrective capabilities to address failures swiftly. Complementing this technological leap is the widespread adoption of SMs, which have transcended their traditional billing role to become pivotal in gathering detailed electricity consumption data [1]. This high-resolution data collection offers invaluable insights into consumer behaviors and lifestyles, underscoring the transformative potential of SMs in understanding and optimizing energy use. SG and SMs embody a comprehensive approach to modernizing the grid, promising a future where energy systems are more resilient, flexible, and attuned to environmental sustainability and the nuanced electricity usage patterns [2].

SM data proves valuable in calculating energy discrepancies, which define the variance between the total energy generated and the energy billed to customers. One can categorize these discrepancies into Technical Losses (TL) and NTL. TL inherently occurs in energy distribution, resulting from the electrical current passing through system components like transformers, sensors, meters, and cables. In contrast, NTL comes from energy consumption that billing does not account for, often due to inaccuracies or fraudulent activities [3]. The growing fraud challenge has led to the development of various detection strategies, increasing NTL. Among these strategies, data-driven approaches have become a prominent solution, leveraging the analysis of customer load profiles and other relevant data to identify discrepancies that indicate NTL. These innovative solutions benefit from cross-disciplinary insights, combining advanced analytics and sector-specific knowledge to improve the accuracy and efficiency of NTL detection.

The distribution of electrical energy incurs significant losses, broadly categorized into TL and NTL, as delineated in various studies [4]. TL are unavoidable and occur naturally during the transmission process, encompassing the dissipation of electricity in its journey through transportation, transformation, distribution phases, and energy measurement activities. These losses are a fundamental aspect of electrical engineering, reflecting current transmission technologies' physical limitations and inefficiencies [5].

On the other hand, NTL refers to electricity consumed but not accounted for in billing. This type of loss can result from a myriad of factors, including illegal connections to the grid, inaccuracies with meter readings due to delayed installation, meter malfunctions, equipment that is either contaminated, defective, or not suited for its intended use, underestimations of actual consumption, improper connections, and oversight of customers' usage [3, 6]. The challenge of NTL extends beyond technical issues, encompassing a wide range of fraudulent activities and operational inefficiencies that lead to significant economic losses for utility companies [7]. Identifying and mitigating NTL remains a critical challenge for the industry, especially in SM (SM) operations. Despite the complexity of the problem and the absence of a one-size-fits-all solution for detecting NTL, ongoing research and the development of innovative methodologies are crucial for tackling this pervasive issue [8, 3]. Effective NTL detection strategies are essential for enhancing power distribution networks' financial sustainability and operational efficiency.

1.2 Motivations and Challenges

NTL is the primary cause of revenue loss in the SG. Electrical utilities incur annual losses amounting to billions due to NTL. Developing countries like India and Brazil each lose 42% and 8% of the total electricity produced annually due to energy theft, respectively [9]. NTL is still as relevant in developed countries as in developing countries. Typically, energy losses in developed countries range from 0.5% to 3% of annual revenues. Though the amount might seem small, financial losses in the United States alone are as high as 6 billion [9].

Although adding categories to a classification problem adds complexity to the classification process, it is crucial to identify the types of fraud committed by end-users.

For example, it can help companies identify the causes of fraud and calculate its financial impact, develop new security measures, and improve the accuracy and efficiency of inspection teams. However, according to Chuwa and Wang [3], existing NTL detection methods cannot effectively identify all types of fraud simultaneously. In addition to being challenging to classify fraud, distinguishing and detecting different types of fraud is more complex using unique Machine Learning (ML) techniques.

In this way, ensemble predictors refer to an ML technique that combines the predictions from multiple models to improve the overall performance of predictive tasks. i.e., aggregated ML models have become increasingly popular after showing excellent results in various research fields by combining the outputs of multiple ML algorithms to achieve better performance than a single classifier. The integration is more heterogeneous when the different time series classifier types are considered to classify these additional parameters and attributes in the user class. In this case, ensemble predictors that consider the time-dependent nature of energy consumption data tend to improve the accuracy and efficiency of detecting fraud, which remains an open question [10, 3].

On another point, it is possible to incorporate an ITQ algorithm that considers the time-series dependence of energy consumption data. Therefore, data orientation in time series can be an essential tool to improve the efficiency of NTL detection as it examines the temporal order of energy consumption data [11]. Data-oriented methods provide better NTL detection because they require less data diversity and incur less infrastructure cost [3]. Furthermore, researchers recognize that a data-centric approach is a promising tool for bridging the cybersecurity gaps that can arise as current fraud mechanisms evolve [12].

The study of energy fraud is an observational science in which one attempts to infer the properties of an unknown system from the analysis of measured time records of its behavior Time Series (TS) [13]. User dynamics analysis, such as fraudulent and non-fraudulent, yields better solutions by understanding the underlying data generation process and identifying distinct patterns. We can apply ITQ to these scenarios by using characterization in NTLs mainly caused by electric fraud. Strategies based on extracting measures such as ITQ (*e.g.*, Shannon entropy, Fisher information, and Statistical Complexity) combined with ordinal pattern methods have yielded relevant progress in distinguishing different TS dynamics [14]. Hence, they depict a promising tool to explain these complex behaviors to improve fraud characterization.

1.3 Research Questions

Based on the Section 1.2, we considered the following research questions for this thesis proposal:

- How to classify non-technical losses in SG systems?
- How to define the type of non-technical losses in SG systems?

- Which is the best classifier to check non-technical losses in SG systems?

1.4 Objectives

This thesis proposal introduces novel methods for the classification and characterization of NTL by leveraging consumer energy consumption data. The classification method utilizes this data, from segmentation to the application of classifiers, incorporating temporal dynamics across preprocessing, training, testing, and validation phases to ensure a thorough time-series analysis. To characterize energy fraud, we employ Information Theory Quantifiers (ITQ) in a structured three-stage process. Initially, we transform the electricity consumption TS into a detailed histogram via the non-parametric Bandt-Pompe (BP) method [15], capturing causal time information. Subsequently, we extract critical metrics from this histogram—Fisher Information (FI) , Statistical Complexity (SC) , and Permutation Entropy (PE) to serve as our analytical framework. These metrics are then plotted on the Complexity-Entropy Causality Plane (CECP) [16] and the FS Plane, where their positions reveal various canonical states indicative of potential fraud. Validated with a comprehensive dataset from the Irish Smart Metering Energy Project [12], which includes over 5,000 private and commercial electricity consumers, our methodology promises to advance the detection and understanding of NTL significantly, offering a strategic advantage in combating energy fraud.

- Develop two distinct algorithms: one for classifying and another for characterizing NTL, employing advanced data analysis methods to identify and distinguish types of NTL in energy consumption data.
- Define the type of NTL for each user: Use the developed algorithms to analyze energy consumption patterns of users, enabling accurate identification of the type of NTL associated with each case, thereby facilitating the implementation of targeted corrective measures.
- Develop and apply these techniques in SM scenarios: Adapt and optimize the algorithms for application on data obtained from SMs, aiming for effective detection of NTL in real-world environments and promoting more efficient energy resource management.

1.5 Contributions

Our main contributions can be summarized in:

- The ensemble considers different Time Series Classifiers (TSC) to create an ensemble predictor, Time Series Forest (TSF) , Catch22, Weasel, k-NN for TS, and Arsenal. After training, the generated predictor classifies new data into thirteen classes (either honest or one of twelve different types of fraud), which facilitates the development of methods to improve the detection of specific types of fraud.

- Summarizing the ITQ contributions: (i) Our work characterizes fraud cases using quantifiers derived from information theory. (ii) We used FI, SC, and permutation entropy to characterize frauds in SG scenarios with little data.
- We evaluate ensemble performance using smart energy data from the Irish Smart Energy Trial [17], which consists of approximately 4700 users with 535 days of sample data. Following the approach of existing work, we add twelve fraud types already defined in the literature to create a synthetic dataset of honest and fraudulent customers and randomly select the amount of fraud data generated among users.

1.6 Text Organization

This text presents the fundamentals of this research based on related works, the main milestones already achieved with the published paper, and the planned advances for future research works. The remaining of this document is structured as follows:

- Chapter 2: Describes the basic concepts of energy losses, frauds, the users used, ML techniques, and ITQ in SG scenarios.
- Chapter 3: Shows the state-of-the-art regarding this thesis proposal. The chapter divides the works into different categories and evaluates them with the proposal.
- Chapter 4: Details all aspects and techniques to evaluate ensemble learning, with all the parameters and metrics used in the classification model.
- Chapter 5: Details all characterization aspects of ITQ techniques and how we performed the evaluation and achieved the results.
- Chapter 6: Concludes the thesis proposal, suggests future works and presents the published works associated with this research.

CHAPTER 2

Basic Concepts

This chapter presents some main concepts about electric power system fraud, ensemble learning, ML techniques, and information theory used within real SG scenarios to identify fraudulent users.

2.1 Smart Meters

SMs represent a significant advancement in digital metering technology, capable of measuring the consumption of electricity, gas, or water in real time or at brief intervals. These devices then communicate this data to the utility provider for efficient monitoring and billing. Contrary to traditional meters, which necessitate manual readings, SMs automate the process of data transmission to the utility company. This automation facilitates more accurate billing and offers consumers and utilities detailed insights into consumption patterns [18]. The primary attributes of SMs include:

1. Real-time Data Collection: SMs are adept at collecting usage data at frequent intervals, potentially as often as every hour or more. This capability ensures a granular understanding of consumer energy or water usage patterns.

2. Remote Communication: Utilizing wireless technology, these meters transmit information directly to the utility provider, obviating the need for manual meter readings and enabling the implementation of dynamic pricing schemes based on time of use.

3. User Insight: Through online platforms or in-home displays, consumers gain access to their usage data, empowering them to identify and exploit energy or water conservation opportunities by adjusting their consumption behaviors.

4. Outage Detection: SMs provide utility providers with immediate alerts regarding power outages and other pertinent issues, enhancing the efficiency of repair and service restoration efforts.

5. Support for Renewable Energy: These meters play a crucial role in integrating renewable energy sources, like solar panels, into the grid by precisely tracking the energy produced and consumed.

The benefits of implementing SMs are manifold:

1. Enhanced Energy Efficiency: SMs promote energy and water conservation among consumers and businesses, potentially yielding cost savings and diminishing environmental impacts.

2. Superior Customer Service: The detailed consumption data furnished by SMs enables utilities to offer personalized advice, dynamic pricing models, and improved customer service, including expedited problem resolution.

3. Optimized Grid Management: For utility providers, SMs are instrumental in improving demand forecasting, load management, and distribution planning, contributing to a more reliable and efficient energy supply.

4. Smart Grid Enablement: As foundational technologies forSGs, SMs facilitate the creation of electricity networks that leverage digital technology to manage and respond to the behavior of all participants (consumers and producers) more effectively.

Challenges and Considerations: Despite the myriad benefits, deploying SMs has its challenges, including concerns regarding privacy, data security, and the potential for unauthorized access. The comprehensive consumption data gathered by SMs can inadvertently expose detailed information about a household's habits and behaviors, underscoring the necessity for stringent data protection measures [1]. Furthermore, the initial investment required for SM deployment and the imperative for public acceptance and trust constitute significant considerations for utility providers. Nonetheless, the global adoption of SMs is on an upward trajectory, propelled by the objectives of enhancing energy efficiency, integrating renewable energy sources, and modernizing utility services overall.

2.2 Ensemble Learning

Ensemble learning represents an advanced ML paradigm that trains multiple models, often called ”weak learners,” to address the same problem and combine them to achieve superior results. The core principle of ensemble learning asserts that by aggregating multiple models, the ensemble achieves higher accuracy and better generalization performance than any single model could independently [18]. This approach effectively reduces errors that variance, bias, or noise in the training data causes. Ensemble learning employs several key methods and techniques, including:

1. Bagging Bagging (Bootstrap Aggregating): This method involves training multiple models on unique subsets of the training dataset, created by randomly sampling the original dataset with replacement. The final prediction usually represents the average of all predictions (in regression problems) or the majority vote (in classification problems). Random Forests exemplify the bagging approach.

2. Boosting: Boosting algorithms sequentially train a series of weak models, with each new model addressing the shortcomings of its predecessors. The strategy emphasizes giving more weight to training instances that previous models in the sequence misclassified. The final model bases its predictions on a weighted sum of the predictions from earlier models. AdaBoost, Gradient Boosting, and XGBoost are examples of boosting algorithms.

3. Stacking (Stacked Generalization): This method trains a new model to combine the predictions of several existing models. After training the original models on the whole dataset, it trains a new model to make the final prediction based on these models' predictions. This second-level model can adjust for the individual base models' biases.

4. Voting: Voting ensembles train multiple models independently and combine their predictions through a simple majority vote (for classification) or averaging (for regression). This method allows using different model types, leveraging the strengths of diverse approaches.

Ensemble methods have proven effective across various domains, including competition platforms like Kaggle, where they frequently win competitions and in real-world applications, from disease prediction to financial forecasting. The main advantage of ensemble learning is its ability to improve prediction accuracy and robustness beyond single models, making it a powerful tool in ML techniques.

2.3 Technical Losses

Technical losses in electrical power systems intrinsically manifest during the transmission and distribution processes due to the inherent inefficiency of electrical components and the physical properties of conductors. Utility companies and system operators express significant concern over these losses because they diminish the overall efficiency of power systems, elevate operating costs, and necessitate increased generation capacities to fulfill consumer demand. This text explores the nature, causes, and mitigation strategies of technical losses, offering a comprehensive overview for academic and industry stakeholders [1].

Primarily, one can categorize technical losses in power systems into transmission and distribution losses. Electricity incurs transmission losses as it travels over long distances from power plants to substations through high-voltage transmission lines. Conversely, distribution losses manifest within the lower voltage networks responsible for distributing electricity from substations to end consumers. The resistance in electrical conductors mainly drives these losses, converting electrical energy into heat in a process described by Joule's law.

1. Resistance in Conductors: The resistance found in transmission and distribution lines emerges as the primary cause of technical losses. These losses, proportional to the current squared flowing through the conductor, become significant in networks with high load demand.

2. Transformer Losses: Transformers play a pivotal role in adjusting voltage levels for efficient transmission and safe distribution, contributing to technical losses. These losses comprise core losses, attributed to the magnetization and demagnetization of the transformer core, and copper losses, stemming from the resistance in the windings.

3. Reactive Power Losses: Necessary for maintaining voltage levels and the operation of certain loads, the transmission of reactive power incurs technical losses. These losses relate to the phase difference between voltage and current in AC systems.

Addressing technical losses is imperative for enhancing the efficiency of power systems and curtailing operational costs. Strategies to mitigate technical losses include:

1. Network Configuration: By optimizing the design and configuration of transmission and distribution networks, one can minimize the distances electricity must travel, thereby reducing losses.

2. Conductor Size and Material: Employing conductors with lower resistance levels, such as high-capacity aluminum or aluminum alloy conductors, facilitates loss reduction. Furthermore, enlarging the cross-sectional area of conductors diminishes resistance.

3. Voltage Level Optimization: Conducting operations at higher voltage levels decreases the current required for the same power transfer, thus reducing conductor resistance losses.

4. Advanced Technologies: The integration of advanced technologies like high-temperature superconductors (HTS) and gas-insulated lines (GIL) can markedly diminish technical losses. However, the high costs of these technologies often limit their usage to selective applications.

5. Reactive Power Compensation: The strategic installation of capacitors and reactors can minimize reactive power flow and its associated losses.

6. Regular Maintenance: Ensuring regular maintenance of electrical equipment, including transformers and conductors, optimizes performance and minimizes losses.

Technical losses, an unavoidable aspect of transmitting and distributing electrical power, significantly improve in efficiency with careful planning, technological innovation, and strategic investment. Reducing technical losses not only enhances power system efficiency but also aids in energy conservation and economic savings. With the growing demand for electricity, the critical focus on addressing technical losses is ever-increasing, underscoring its importance in the energy sector.

2.4 Frauds

Frauds within the electrical system encompass a spectrum of illicit activities that culminate in the unauthorized utilization or distribution of electricity. These activities span from meter tampering, aimed at underreporting electricity usage, to intricate strategies devised to circumvent or manipulate billing systems [3]. The ramifications of such frauds are profound, precipitating financial deficits for utility corporations, escalated costs for law-abiding consumers, and safety hazards due to improperly established electrical connections and usage. The spectrum of fraud in the electrical system encompasses, but is not limited to:

1. Meter Tampering: Individuals may engage in the physical alteration of electricity meters or their connections, employ magnets to decelerate the recording mechanism, or infiltrate SMs to modify digital records, thereby reducing the reported energy consumption.

2. Bypassing Meters: Certain individuals opt for direct connections to the power grid, bypassing meters altogether. This action results in electricity usage that remains unrecorded by the meter, rendering the utility company unable to charge for the actual consumption.

3. Billing Fraud: This category involves altering billing records or account details to decrease the amount payable. Tactics include falsifying meter readings, transferring debts to another account, or establishing phantom accounts with lower tariffs.

4. Illegal Resale: Offenders may illicitly connect to the electrical grid and resell access to electricity at discounted rates to neighbors or businesses, thereby accruing profits without remunerating the utility company for the electricity consumed.

5. Theft of Services: This broad classification encompasses unauthorized access to electrical services, such as tapping into streetlights, public utility connections, or other power sources without authorization.

Addressing these frauds necessitates a multifaceted approach comprising technological interventions, legislative actions, and initiatives to raise public consciousness. The deployment of SMs, capable of real-time electricity usage reporting and detecting anomalous consumption patterns, has posed challenges for individuals attempting meter tampering or bypassing. Furthermore, utility corporations increasingly resort to sophisticated software for analyzing consumption patterns to pinpoint potential fraud. Legislative actions have intensified penalties for electricity theft and tampering. Concurrently, campaigns to heighten public awareness aim to illuminate electricity fraud's financial and safety implications, including the risk of electrical fires and other hazards. In summation, fraud in the electrical system represents a obstacle for both utility providers and consumers, necessitating relentless endeavors to unearth, thwart, and penalize illicit electricity consumption practices.

2.5 Machine Learning Classifiers

ML classifiers categorize data into distinct classes or categories based on data features. These algorithms form a fundamental part of supervised learning, aiming to learn a model from labeled training data to predict unseen data. We can use classifiers for various applications, including spam detection, image recognition, medical diagnosis, and more [19]. This thesis utilizes several ML classifiers:

2.5.1 Signature

The (generalized) signature method results from a unifying framework that collects different feature extraction techniques for multivariate time series equations. The collection employs augmentations, windows, transformation, and rescaling, all grouped in a single mathematical framework.

2.5.2 22 Canonical Time-series Characteristics

22 Canonical Time-series Characteristics (catch22) is a dynamic and commonly used technique for time series data. Catch22 captures time series' diverse and interpretable characteristics according to their properties, including linear and nonlinear autocorrelation, continuous differences, value distributions, outliers, and volatility scaling properties [20]. This technique utilizes a reduction in dimensionality from 4791 to 22, correlates with a roughly 1000-fold decrease in computation time, and scales almost linearly with time series length, despite an average 7% reduction in classification accuracy.

2.5.3 Weasel

The word extraction for time-series classification is a TSC method that is scalable and accurate. Weasel has considered the differences between classes in the feature discretization process rather than relying on fixed, data-independent intervals; this results in a highly discriminatory feature set. Weasel does not treat each fixed-length window as an independent feature but uses windows of different lengths and considers the windows' order. Weasel applies aggressive statistical feature selection instead of simply using all features for classification; this results in smaller function space and dramatically reduces runtime without sacrificing accuracy [21].

2.5.4 Time Series Forest

The tree-ensemble classifier, Time Series Forest (TSF), introduces a novel measure known as Entrance (Entropy and distance) gain for identifying high-quality splits, incorporating entropy gain and employing two One-nearest-neighbor algorithms with dynamic time warping (Xi et al., 2006). TSF adopts a random feature sampling strategy, achieving linear computational complexity relative to the time series length. TSF proposes addressing challenges with two main strategies. First, it introduces a new splitting criterion, the Entrance gain, which merges the entropy gain with a distance measure for pinpointing high-quality splits. Experimental studies across 45 benchmark datasets have demonstrated that the Entrance gain significantly enhances TSF's accuracy. Second, by randomly sampling features, TSF maintains linear computational complexity about the time series length. Moreover, TSF independently grows each tree, allowing modern parallel computing techniques to accelerate the process [22].

2.5.5 k-Nearest Neighbour for Time Series

The importance of interpretation and insight makes k-Nearest Neighbour (k-NN) methods stand out in data analysis. Because k-NN methods provide transparency, they produce interpretable models. This transparency is also crucial in time series analysis, where comparing time series side by side can highlight similarities and differences between methodologies. However, using k-NN methods for time series analysis introduces additional challenges in developing metrics that accurately capture the similarity between time series. Even if one time series stretches or shifts relative to another, the two can still be similar. The similarity might also rely on short or even tiny signatures within the time series[23].

2.5.6 Support Vector Machine

Support Vector Machine (SVM) is a classic supervised ML technique that uses a convex optimization algorithm to maximize the distance (margins) between two categories.

In other words, it tries to predict the correct category for each data successfully.

2.5.7 Extreme Gradient Boosting

Extreme Gradient (XGBoost) Boosting is one of the most important ensembles that use gradient boosting techniques. Such a boosting technique increases the influence of high-performing models sequentially using successive weak learners, and the gradient tries to minimize errors in these sequential models. XGBoost goes further with this strategy by employing level-wise (depth-first) parallel trees gradient boosting processing, tree-pruning, and regularization method to avoid overfitting.

2.5.8 CatBoost

Catboost uses the gradient boosting method, which works well with categorical features by creating symmetrical decision trees using permutation. Therefore, it is ideal for data from different sources. Its ensemble enables good results with few preparations (parameter tuning) and few Algorithm runs, in opposition to long-running deep learning models.

2.5.9 LightGBM

Light Gradient Boosting Machine (also known as LightBoost) uses the gradient boosting method on decision trees by employing leaf-wise (best-first) decision, focusing only on the leaf to carry the maximum gain and minimum loss. Therefore, with this implementation, the LightBoost technique was designed to be more efficient and quicker using a larger gradient during training tests.

2.5.10 Arsenal

An ensemble of Rocket transformers employs a Ridge classifier with built-in cross-validation. Rocket (Random Convolutional Kernel Transform) transforms time series using numerous random convolution kernels, including kernels with random lengths, weights, warping, dilation, and padding. A linear classifier then learns from the transformed features. The combination of Rocket and logistic regression creates a single-layer convolutional neural network with random kernel weights, with the transformed features feeding into the trained softmax layer. However, this ensemble applies a ridge regression classifier to all but the largest datasets, offering the advantage of rapid cross-validation on the regularization hyperparameters without affecting other hyperparameters. However, logistic regression, trained with stochastic gradient descent, scales better. For extensive datasets, logistic regression becomes the method of choice when the training sample size significantly exceeds the number of features. [24].

Each of these classifiers has its strengths and weaknesses and is chosen based on the specific characteristics of the data and the task at hand. The choice of classifier can depend on several factors, including the size and dimensionality of the dataset, the linearity of the decision boundary, and the computational efficiency required.

Conventional ML models often perform sub-optimally because they present high bias, such as a low degree of freedom classifiers or high variance [25]. Thus, in ensemble learning, multiple predictors (often called ”weak learners” or ”basic models”) are trained to solve the same problem [26]. An ensemble predictor builds a more robust predictor with better results than single-classifier predictors. Hence, an ensemble predictor reduces the bias and/or variance of basic classifiers by combining several to create an aggregated learner (or ensemble predictor) that achieves better performances [19].

We consider a voting class classifier for the ensemble predictor, combining different ML classifiers conceptually. The voting classifier accounts for every classifier’s vote for each returned class. As a result, the final classification can be decided either by a majority vote (hard voting) or by the mean value of the individual probabilities produced by every classifier (soft voting). In soft voting, each classifier evaluates a sample’s matching probability for every class. The sum of all probabilities will always be 1. After all, classifiers calculate the probabilities, and soft voting finds the weighted mean of each class’s probability and establishes the class’s final prediction with the highest mean value. It might apply weight to prioritize a particular classifier over others.

2.6 Information Theory Quantifiers

Information theory is a branch of applied mathematics and electrical engineering involving the quantification of information. Developed by Claude Shannon [27] in the mid-20th century, it provides a mathematical framework for understanding information transmission, processing, and storage. Central to information theory are several key quantifiers that measure different aspects of information. Here are the primary concepts:

1. Entropy (H): Entropy measures the unpredictability or uncertainty of a random variable. It quantifies the average amount of information produced by a stochastic data source. For a discrete random variable X with possible values $x1, x2, \ldots, xn$ and probability mass function $P(X)$, the entropy $H(X)$ is defined as:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_b P(x_i) \tag{2.1}$$

   Where b is the base of the logarithm used, in information theory, b is often 2 (bits) but can also be e (nats) or 10 (digits).

2. Conditional Entropy (H(X—Y)): Conditional entropy measures the average amount

of information needed to describe the outcome of a random variable X given that the value of another random variable Y is known. It quantifies the uncertainty remaining about X when Y is known.

3. Joint Entropy (H(X, Y)): Joint entropy of two random variables X and Y measures the uncertainty associated with the joint distribution of X and Y. It quantifies the total information needed to describe X and Y simultaneously.

4. Mutual Information (I(X; Y)): Mutual information measures the amount of information one random variable contains about another. It quantifies the reduction in uncertainty about X due to the knowledge of Y and is symmetric, meaning that $I(X;Y) = I(Y;X)$. It is defined as the difference between the entropy of X and the conditional entropy of X given Y:

$$I(X;Y) = H(X) - H(XY) \tag{2.2}$$

5. Kullback-Leibler Divergence ($D_{KL}(P||Q)$): Also known as relative entropy, the Kullback-Leibler divergence is a measure of how one probability distribution P diverges from a second, expected probability distribution Q. Though it is often referred to as a distance measure, it is not symmetric and does not satisfy the triangle inequality. It is defined for discrete variables as:

$$D_{KL}(P\|Q) = \sum_i P\left(x_i\right) \log \frac{P\left(x_i\right)}{Q\left(x_i\right)} \tag{2.3}$$

6. Cross Entropy: Cross entropy measures the average number of bits needed to encode data from one distribution using the optimal code for another distribution. ML often uses it to define the loss function for classification problems.

These quantifiers are foundational in various fields, including communications, data compression, ML, and cryptography, helping to optimize processes and algorithms by quantifying how much and how efficiently information is encoded, transmitted, and decoded.

2.7 Chapter Conclusions

This chapter described the main basic concepts of the theoretical approach proposal. Given the complexity of the scenario, we can improve fraud detection and data handling in the electric power system and achieve deception with high reliability and security with low data usage.

CHAPTER 3

Related Works

This chapter presents the state-of-the-art regarding fraud, NTL scenarios, and algorithms we use to identify or classify NTL in SG scenarios, such as ML and ITQ models, algorithms, and techniques. Our proposal approaches some influences and similar concepts as some of the related works but also considers other innovative questions.

3.1 Related Works

Gunturi et al. [1] proposed an energy theft detector based on ensemble classifiers, which uses real-time data from SMs to analyze trends in user consumption behavior. The proposal uses the Synthetic Minority Over-sampling Technique (SMOTE) to balance the minority class data and generate synthetic fraud data using the same work fraud case. The authors used six ensemble algorithms to classify users, including ADaptive Boosting (ADB) , CATegorical boosting (CAT) , extreme gradient Boosting (XGB) , Light Gradient Boosting (LGB) , Random Forest(RF) , and Extra Trees (ET) . They used an ensemble machine learning (ML) classifier-based energy theft detector that uses real-time **SM** data to study consumer usage behavior trends—the proposed model for finding energy theft in SGs. Theft can occur at any level of the system. In this work, we assume that SMs are installed across all consumers. It did not consider the time series of the data when separating the training and testing. In addition, this work did not use any time series classifier to compare with their work.

Passos Júnior et al. [28] propose evaluating optimum-path forest (OPF) clustering for non-technical losses. Utilize two private datasets a Brazilian electrical power company provided: one composed of commercial profiles and another managed to find industrial consumers. Another main contribution of this paper is to model the problem of non-technical loss identification as an anomaly detection task. The classifier is trained with regular consumers only.

Barja-Martinez et al. [29] proposed a holistic analysis of AI applications in distribution power systems after identifying and classifying the different data-driven techniques for power systems and the data sources involved in the data acquisition. These applications include operation and monitoring, predictive maintenance, non-technical loss detection, forecasting, flexibility management, and planning of distribution grids.

Jokar, Arianpoo, and Leung [11] proposed a consumption pattern-based energy theft detector. This detector uses a Support Vector Machine (SVM) classifier to analyze each user's samples and classify them as honest or fraudulent. However, the results obtained differed from other ML algorithms that performed better. Punmiya and Choe [30] consider three state-of-the-art gradient boosting algorithms, namely XGBoost, catBoost, and LightBoost, for NTL detection. The detector shows significantly better false positive rate performance compared to related methods. However, the authors did not use techniques to select hyperparameters for the classifier, nor did they use cross-validation techniques to split the data.

Bastos et al. [31] proposed a data-oriented ensemble predictor based on time series classifiers (TSC) for NTL detection called DETECT. The proposed predictor is time-series-oriented, from how the data is split to implementing the classifiers. DETECT considers an ensemble predictor of five TSC algorithms, namely, Time Series Forest (TSF), Residual Network (ResNet) , Inception Time, Time Le-Net (t-LeNet) , and Multi-Channel Deep Convolutional Neural Network (MCDCNN) . However, this work does not deal with all fraud models currently cataloged in the literature and does not perform well for different types of fraud, in addition to not worrying about data processing time.

Jindal et al. [32] developed a method to identify NTLs in the grid's transmission and distribution levels. At the distribution level, the identification process consists of a DT algorithm paired with a Support vector machines (SVM) algorithm. The SVM classifies users according to their consumption patterns, while the Decision Tree (DT) estimates the energy consumption for each user based on the temperature, number of people living in residence, number of electrical appliances, time of the day, season, and so on. However, privacy concerns arise from its extensive use of user data as input to the SVM.

Punmiya and Choe [30] proposed an exhaustive analysis comparing the three most recent Gradient Boosting Classifiers (GBCs) : Extreme Gradient Boosting (XGBoost) , Categorical Boosting (CatBoost) , and Light Gradient Boosting Method (LightGBM) . This study aims to develop a Gradient Boosting Theft Detector (GBTD) that incorporates these GBCs, featuring a preprocessing module designed through feature engineering to enhance the detection rate (DR) , lower the false positive rate (FPR) , and optimize time complexity. Within the GBTD classifiers, the preprocessing component includes a stochastic feature generation function that betters the FPR and the DR by leveraging combinations of daily electricity consumption figures as attributes. Moreover, this module incorporates a feature extraction mechanism employing Weighted Feature Importance (WFI), which significantly diminishes training time complexity by eliminating non-essential attributes (noise) from the consumer's data set, thereby also contributing to a reduction in storage requirements for consumer information in SG.

Messinis, Rigas, and Hatziargyriou [33] proposed a hybrid method for NTL detection composed of three blocks that can include more types of fraud. The first block, or module, performs feature extraction from time-series consumption data and trains an SVM classifier to identify NTL. The second module calculates network self-sensibility tension based on meter data and compares it to theoretical values obtained with the network topology. The third module solves an optimization problem whose objective is to minimize losses. Although this method is hybrid and novel, it requires data that is often unavailable, as was the case in this study, which involved generating the data used for validation.

Souza et al. [34] proposed a method to detect and identify electrical energy theft in distribution systems. This method integrates a system of static state estimation (SSE) and phasor measurement unit (PMU) techniques at the beginning and end of the feeders. They compare the SSE results with the total consumption reported by SMs. If the discrepancy exceeds 10%, they scrutinize all customers connected to that feeder, classifying them as honest or fraudulent. To facilitate this, they employed Self-Organizing Maps (SOM) to group users with similar consumption patterns and used a Multilayer Perceptron Artificial Neural Network (MP-ANN) for the final user classification. Although this method yielded satisfactory outcomes, it could have benefited from a more robust data validation technique, such as cross-validation, and did not explore the use of Time Series Classification (TSC).

Aoufi et al. [35] proposed the necessity of detecting energy theft through the precise identification of SM data manipulations, emphasizing the critical role of data integrity in power dispatching and dynamic pricing. They further highlight the importance of swift detection times, given the real-time transmission of SM data. They acknowledge that while statistical models, such as ARIMA, offer advantages in execution speed over machine learning and deep learning models, their accuracy may need to be improved. Building on this insight, we introduce a hybrid energy theft detection system that integrates one forecasting-based statistical model with two forecasting-based deep learning models to enhance the accuracy and efficiency of energy theft detection.

Zheng et al. [8] introduced an electricity theft detection system that processes each piece of consumption data through two components: (a) a Wide CNN, featuring a fully connected layer, and (b) a Deep CNN, comprising multiple convolutional layers. They combine the outputs of these two components using the Sigmoid function to determine whether the consumption values indicate normal usage or an attack.

Ahir and Chakraborty [36] introduced a pattern-based and context-aware methodology for detecting energy theft, distinguishing multiple electricity usage patterns for each user according to the calendar context, including weekdays, weekends, seasonal variations, holidays, and specific blocks of hours within a day. They combined the k-NN algorithm and Dynamic Time Warping (DTW) to uncover the relationships between different consumption patterns. They tested their approach using a dataset provided by the Calcutta Electric Supply Corporation (CESC).

Hasan et al. [37] proposed a binary classification model that integrates both Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models. They trained and tested this model using the SGCC dataset, which features significantly fewer malicious data instances than normal ones. To counteract this imbalance and prevent biased classification, they applied the Synthetic Minority Over-sampling Technique (SMOTE).

Adeli et al. [38] proposed using an optimization algorithm to fine-tune the parameters of detectors. This algorithm aims to enhance the performance of SMO-based attack detectors by minimizing the discrepancies between actual attacks and estimated attack signals, reducing attack detection times, and improving the accuracy of attack detection. Adjusting detector parameters considers the possibility of unknown attacks, treating the attack vectors as random variables within the optimization process. As a result, the nature of the optimization problem evolves. To navigate this complexity, Adeli employs a differential evolution algorithm for the dynamic adjustment of parameters in the SMO-based attack detection method. It is important to note that, in contrast to learning-based attack detection algorithms, which adjust parameters in a real-time or online manner, the parameter adjustment for the SMO-based attack detector occurs offline. This offline adjustment enables its application in real-time attack detection scenarios.

Zidi et al. [39] proposed an approach for automating theft detection by applying data analysis techniques. This method leverages energy consumption data from various consumers, applying and comparing multiple machine learning techniques to identify and detect abnormal consumption behaviors. Zidi and their team designed an effective theft generator to facilitate the analysis of energy consumption behavior in SG environments. They introduced a multi-class theft detection dataset to evaluate classifier performance and serve as a benchmark. The team developed an intelligent autonomous detection system capable of identifying six distinct types of theft. Furthermore, they conducted extensive simulations to characterize the performance of five different machine learning techniques: K-Nearest Neighbors (KNN), Decision Trees (DT), Random Forest (RF), Bagging, and Artificial Neural Networks (ANN).

Xia et al. [40] proposed a hybrid method enhanced by innovative concepts to refine the model to better align with real-world power grid conditions and augment detection accuracy. This approach involves focal loss during model training to amplify the influence of a limited number of samples on model optimization. Furthermore, Xia implemented a channel-dimensional adaptive attention module to intelligently combine the feature expressions derived from both broad and deep Convolutional Neural Networks (CNNs), thereby improving the precision of model training.

3.2 Chapter Conclusions

Table 1 summarizes the main characteristics of previous works aimed at NTL detection in terms of the split for time series, ML algorithms employed, ensemble predictor considered, data used for NTL prediction, TSC, several output classes for NTL prediction, and characterization based on ITQ. Another contribution of this thesis is the formation of a whole. Firstly, the ensemble is more heterogeneous, connecting different TSC types for the classification of different classes to classify the parameters and properties found in the user class. Secondly, it shows that a single classifier has many application dependencies, as they are limited to classifying binary files and cannot accurately classify different classes, and also do not perform characterization of users with similar characteristics. NTL detection remains a challenge, and to our knowledge, no literature studies characterize fraud from the user's energy consumption based on heterogeneous classifier and ITQ. In this context, developing methods that compare different ML algorithms and maintain the temporal dependence of the data is paramount in providing results that better reflect real-life scenarios.

Table 1: Summary of Related Work.

Works	**Split for Time Series**	**Techniques**	**Ensemble Predictor**	**Data Type**	**TSC**	**Number Of Classes**
Gunturi et al. [1]	No	RF, ET, CAT XGB, ADB, LGB	Yes	Electricity Consumption	No	2
Passos Júnior et al. [28]	No	OPF, k-Means, GMM, AP, Birch, SVM	No	Electricity Consumption	No	2
Barja-Martinez et al. [29]	No	Correlation, DT, LR, LogR, MLP, DNN	No	Operational Data, Weather Data, Electricity Consumption, Social Media, GIS, Customer behavior data	No	2
Jindal et al. [32]	No	SVM, Decision Tree	No	Electricity Consumption	No	2
Bastos et al. [31]	Yes	SVM, CatBoost, XGBoost, LightBoost, TSF, ResNet, Inception, TLENERT, MCDCNN and Detect	Yes	Electricity Consumption	Yes	7
Jokar, Arianpoo and Leung. [11]	No	SVM	No	Electricity Consumption	No	2
Punmiya and Choe [30]	No	XGBoost, CatBoost, LightBoost	No	Electricity Consumption	No	2
Messinis, Rigas and Hatz. [33]	No	SVM	No	Electricity Consumption	Yes	2
Souza [34]	No	SSE, PMU, SOM and MPANN	No	Electricity Consumption	Yes	2
Aoufi et al. [35]	Yes	ARIMA	No	Electricity Consumption	Yes	2
Zheng et al. [8]	Yes	CNN and Deep CNN	No	Electricity Consumption	Yes	2
Ahir and Chakraborty [36]	Yes	K-NN and DTW	No	Electricity Consumption	Yes	2
Hasan et al. [37]	No	XGBoost, CatBoost, LightBoost	No	Electricity Consumption	No	2
Adeli et al. [38]	Yes	SMO-based attack	No	Electricity Consumption	No	2
Zidi et al. [39]	Yes	KNN, DT, RF, RF ANN	Yes	Electricity Consumption	Yes	6
Xia et al. [40]	No	CNN	No	Electricity Consumption	No	2
The Algorithm	**Yes**	**Signature, CatBoost, XGBoost, LightBGM, TSF, Weasel, ROCKET, SVM, Catch22, k-NN and HybridForest**	**Yes**	**Electricity Consumption**	**Yes**	**13**

CHAPTER 4

Classifier of Fraud Detection and Ensemble Machine Learning Mode for NTL in Data-oriented

This chapter presents the classification of frauds based on ensemble learning, the results achieved with the classifier evaluation, the techniques chosen, the amount of data used, and the conclusions we obtained. The ensemble is composed of ML techniques that can better define users for each type of fraud, thus having better results for a more heterogeneous dataset, resulting in a scenario closer to reality where users are different. From this, we demonstrate our results and verify other more straightforward techniques, considering a more realistic model.

4.1 Performance Evaluation

This section describes the methodology and performance metrics used to evaluate the predictors for NTL detection. We compared the performance results obtained with the Algorithm with other TSC and non-TSC ML algorithms. We used the following performance metrics: Precision, F1-score, Accuracy, False Positive Rate (FPR), and Recall to evaluate the effectiveness of the Ensemble algorithm.

Figure 1 presents an overview of our Ensemble Algorithm, which considers user energy consumption as the data input for classification. The Algorithm relies on an ensemble predictor of five of the most recent TSC classifiers. Once trained, the resulting predictor outputs the category to which each sample belongs, either benign or one of the twelve types of fraud. The Algorithm consists of three main functions: i) generation of fraud data, ii) data pre-processing, and iii) training using the ensemble predictor based on TSC.

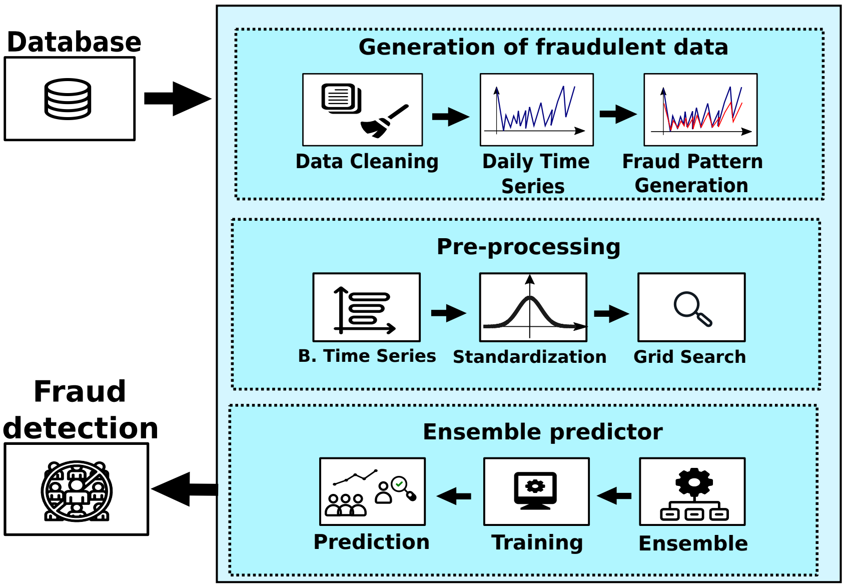

Figure 1: Overview

Finding public records with honest and fraudulent user labels for data classification is difficult. In this sense, we use consumption data obtained in continuously monitored scenarios. Participants accepted the terms of commitment to install SM in their homes. It is a reasonable assumption that all samples belong to honest users [41]. To cope with this issue, we consider the Irish smart energy dataset [17], a widely used dataset for SG scenarios. The Irish dataset contains the consumption data of 4710 domestic and commercial clients and 535 days of readings collected in Ireland between 2009 and 2010, enabling validation of the fraud detection method in a large-scale scenario. This dataset considers Intelligent Electronic Devices (IEDs) registering consumption readings every half hour, where the first reading corresponds to the interval between 0h0min0s and 0h 29min 59s, and the second reading corresponds to the interval between 0h 30min 0s and 0h 59min 59s and so forth. Thus, every day is composed of 48 sequential readings, *i.e.*, , sample n of vector EC equals 48.

4.1.1 Generation of Fraudulent Data

The *generation of fraud data phase* starts with the load of the vector with real daily consumption readings. Afterward, we need to clean the data to find errors or null values in the dataset, detect outliers, and transform them into acceptable types using scale-changer methods [1]. For instance, the number of samples must be the same for each daily vector EC since some classifiers can not handle different length time series. Hence, we drop daily vectors without the expected samples, as those could represent failures in the SG, meaning non-reliable information. This step represents a reduction of only 1% in the original number of daily vectors and does not represent a significant loss of information.

The organization of the Irish dataset features four columns: the id column for unique user identification within the dataset, the day column to mark the date of each measurement, the measurement column to denote each of the 48 daily samples, and the consumption column to log the energy consumption for each day. To facilitate fraud detection, it becomes necessary to reorganize these readings into daily vectors. In this structure, each row corresponds to a day's worth of readings, with columns reflecting the consumption values for each measured interval. This reorganization proves essential for accurately identifying fraud cases, as it aligns the dataset structure with the requirement for analyzing daily consumption patterns to detect fraudulent activities. Specifically, each row contains a time series with samples ranging from 0 to 47, the class label (*i.e.*, , either zero for regular consumption or a number between 1 and 12 for fraudulent patterns), the day this data, and the corresponding user.

The dataset exclusively comprises honest consumption data, given its generation from continuous monitoring of customers who had consented to the terms of the agreement and completed questionnaires before and after the measurement period. Similar to other work using the Irish dataset, we also add different types of fraud to generate a synthetic dataset with honest and fraudulent customers due to the lack of real fraud samples in the dataset [3]. Specifically, we add twelve fraud types already defined in the literature for the whole time series, from the first to the last reading, as other works [30, 34, 1] have also done. In general, frauds report less energy than the actual energy consumed or redistribute the energy consumption at different times to take advantage of the varying billing system [1]. It is important to mention that the amount of samples for each fraud in the dataset is smaller because those classes would be less frequent than the normal class on a real problem. For instance, the abnormal classes have 25% fewer samples than the regular class.

These frauds have the same objective of reducing the overall electricity bill. As attackers have different motives, behaviors, and random energy consumption, producing NTL frauds that include all behavior of malicious patterns is challenging. This section discusses various models of synthesized NTL frauds suggested by current works.

1. **Fraud:** We multiply all the readings x_t (electricity consumption in kW per hour) with the same real pseudo-random number α in a predetermined interval between min and max values as Eq. (4.1) and Figure 2 show, which can be considered as the most traditional observed fraud on the SM scenario [9]. Specifically, α values closer to min mean the higher the severity of the fraud. It is a fraud pattern in which the user artificially reduces daily consumption continuously (e.g., α value of 0.3 means that SM reports only 30% of energy consumption) and by the same proportion between daily measurements [9]. fraud can be called anormal consumption, because do not follow the correct consumption.

$$f_1(x_t) = x_t * \alpha, \ \ (where \ \ \alpha = random.uniform(min, max)) \tag{4.1}$$

2. **Fraud:** We multiply each meter reading x_t with a different integer random number

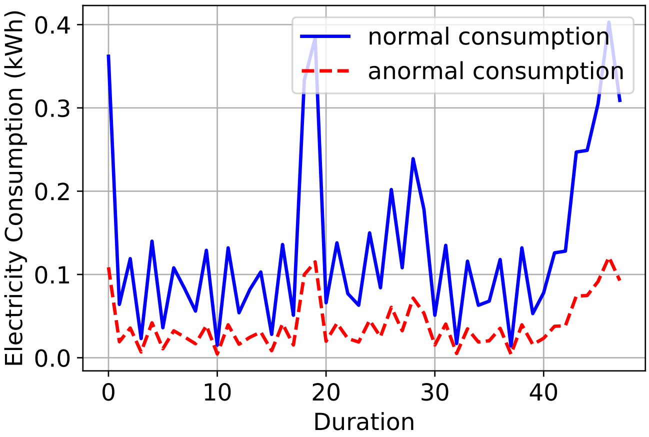

Figure 2: Fraud 1

$\gamma_t \in [0, 1]$, as Eq. (4.2) and Figure 3 show, also known as an on-off fraud. This fraud type might indicate the malfunction of an SM or grid infrastructure. It is a fraud pattern in which readings are either interrupted or canceled for certain times of the day, i.e., it replaces the consumption samples for zero each day in a random duration. This fraud can be easily detected, especially after a long period of zero reporting [9].

$$f_2(x_t) = x_t * \gamma_t, \quad (where \ \ \gamma_t = randint.uniform(0, 1)) \tag{4.2}$$

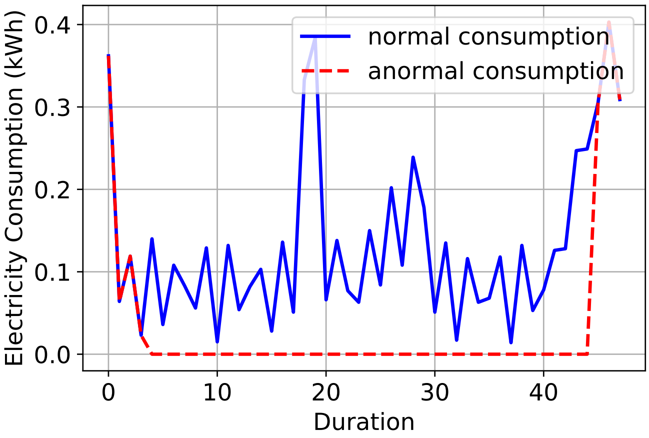

Figure 3: Fraud 2

3. **Fraud:** We multiply each meter reading x_t by a different real pseudo-random number β_t between min and max values in a predetermined interval, as Eq. (4.3) and Figure 4 show. Specifically, the fraud level increases as the βt value decrease. This fraud has a similar equation compared to fraud 1, but the βt value is different for each sample reported. It means the NTL might not occur continuously, and there may be some discontinuous reporting of "fraudulent" values. Therefore, the user

produces different reduction rates across measurements in this fraud pattern.

$$f_3(x_t) = x_t * \beta_t, \ \ (where \ \ \beta_t = random.uniform(min, max)) \qquad (4.3)$$

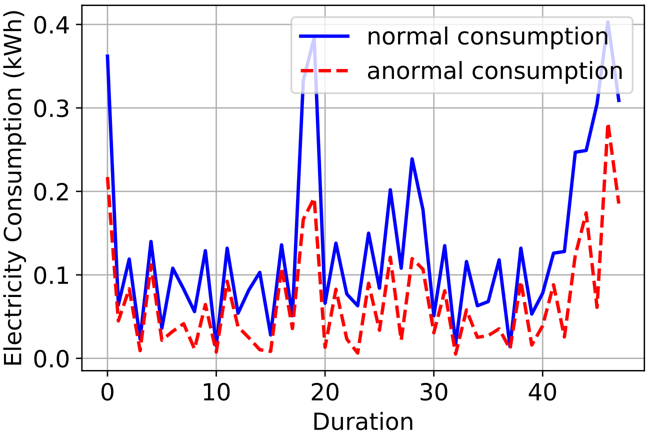

Figure 4: Fraud 3

4. **Fraud:** We use the mean values of the readings multiplied by areal pseudo-random number θ_t in a predetermined interval between min and max values, as Eq. (4.4) and Figure 5 show. It represents the average of the readings made over the day with a continuous reporting of "fraudulent" values. In this fraud pattern, the user falsifies the trend and daily consumption, producing a new pattern completely different from the original one.

$$f_5\left(x_t\right) = \text{mean}(x) \times \theta_t, (\text{ where } \theta_t = \text{random}(\text{min}, \text{max})) \qquad (4.4)$$

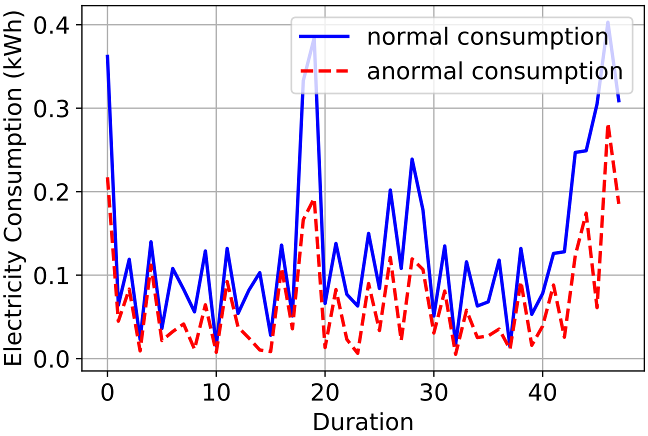

Figure 5: Fraud 4

5. **Fraud:** We use the mean values of the readings for all measured samples, as Eq. (4.5) and Figure 6 show, representing the exact average of readings over the day.

This fraud can be easily detected by analyzing the consumption profile because a constant consumption does not reflect the average electricity consumption that needs to change randomly over time [9].

$$f_5\left(x_t\right) = \text{mean}(x_t) \tag{4.5}$$

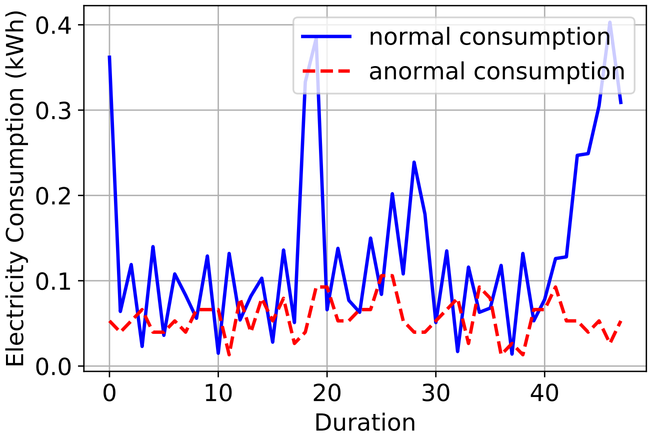

Figure 6: Fraud 5

6. **Fraud:** Eq. (4.6) and Figure 7 describe a pattern in which the readings have their chronological order changed, where it does not steal electricity but by shifting high consumption from peak to off-peak [9]. It is worth noting that frauds 4.5 and 4.6 describe fraud patterns of target systems that bill clients differently according to the time of the day.

$$f_6\left(x_t\right) = x_{T-t} \text{ (where } T \text{ is the sample size per day)} \tag{4.6}$$

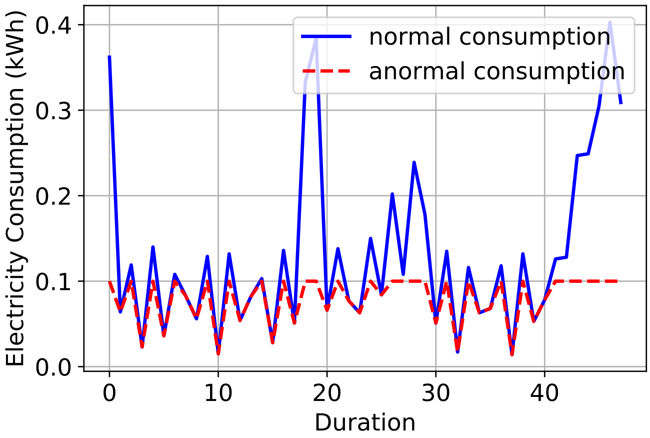

Figure 7: Fraud 6

7. **Fraud:** This fraud simulates an energy theft that takes place over a certain period [28, 42]. The duration of the fraud is selected randomly, and all measurements in

this duration are reduced. The start time and duration of the frauds vary according to the attackers' needs. Yet, most attackers reduce their electricity during peak hours, days, and seasons. Figure 8 shows the variation of fraud 4 and is formulated by

$$f_4(x_t) = \gamma_t * x_t, \gamma_t = \begin{cases} \alpha, & t_s < t < t_x \\ 1, & \text{else} \end{cases} \tag{4.7}$$

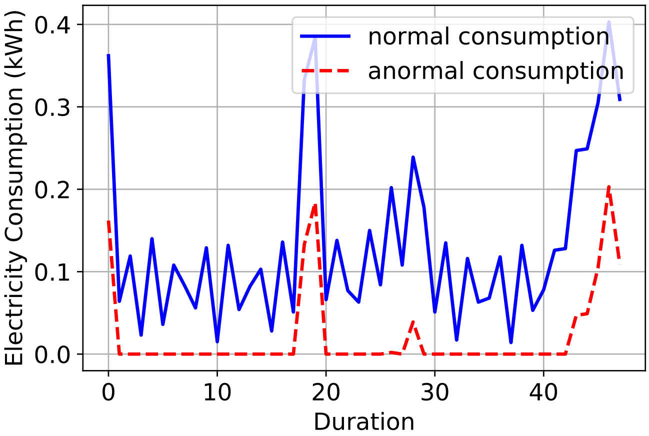

Figure 8: Fraud 7

8. **Fraud:** A random cut-off point (often arises within the context of statistical analysis, ML algorithms, and decision trees. It refers to a method of dividing data into subsets based on a threshold value that is chosen randomly. In ensemble learning, particularly with methods that involve bagging (Bootstrap Aggregating) or boosting, random cut-off points can contribute to creating a diverse set of models. Each model in the ensemble might use different cut-off points for the same features, leading to varied decision boundaries. This diversity is beneficial because it allows the ensemble to cover a broader range of data patterns, making the combined model more robust and accurate than any single constituent model.) is selected, and all values of energy consumption above that point are replaced by a cut-off value [43]. For this type of fraud, the attacker does not want to exceed the maximum predefined limit. Consequently, all consumption exceeding the maximum allowable limit is reduced. The cut-off point should be carefully selected, as a very low cut-off point may cause the meter to report a series of constant consumption that can be easily detected. Also, if the cut-off point is too high, the stolen amount will be too low to benefit the attacker. This fraud is formulated in Eq. 4.8 and Figure 9. Where, α is cutoff point and $\alpha < max(x_t)$.

$$f_6(x_t) = \begin{cases} x_t, & x_t\alpha \\ \alpha, & x_t > \alpha \end{cases} \tag{4.8}$$

9. **Fraud:** In this fraud, a random cut-off point is selected and subtracted from the

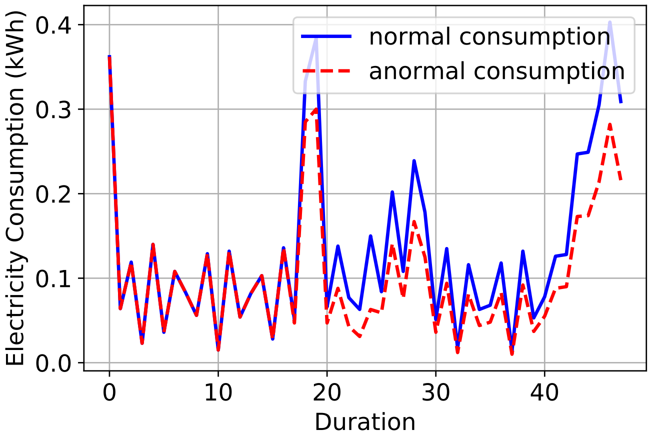

Figure 9: Fraud 8

actual sample [43], as shown in Eq. 4.9 and Figure 10—example of daily variation of fraud 9. If the result obtained is less than zero, zero is reported. This fraud may be due to injecting false data or bypassing the meter by connecting the load directly to the distribution transformer. Thus, the amount consumed by the load is not recorded by the meter.

$$f_7(x_t) = \max(x_t - \alpha, 0) \tag{4.9}$$

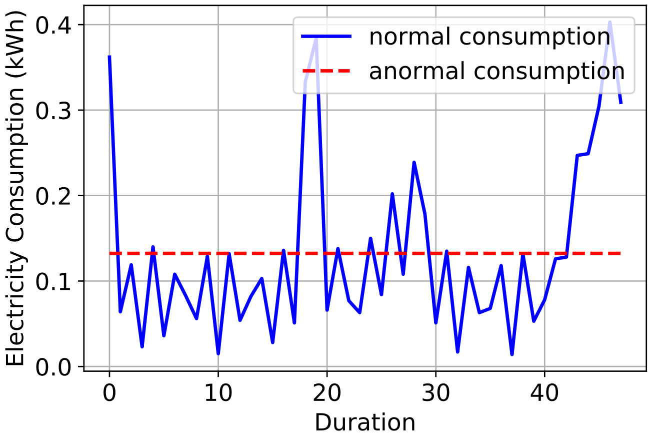

Figure 10: Fraud 9

10. **Fraud:** The authors of [44] proposed this fraud and called it a smart fraud. It is smart because energy consumption does not decrease abruptly but gradually decreases until the fraud reaches its maximum intensity. It stays at this point for the rest of the fraud duration. The gradual decrease is determined by the rate of change in the intensity of the fraud. Eq. 4.10 a mathematical model, and Figure 11 shows the variation of fraud 10, respectively. Where it is attacked intensity $(0 < i_t < 1)$, s, indicates the rate of change in fraud intensity, and t max is the time

with maximum intensity.

$$f_8(x_t) = (1 - i_t)\, x_t, i_t = \begin{cases} i_{\max}, tt_{\max} \\ s\,(t - t_s)\,, t_s < t < t_{\max} \\ 0, t < t_s \end{cases} \tag{4.10}$$

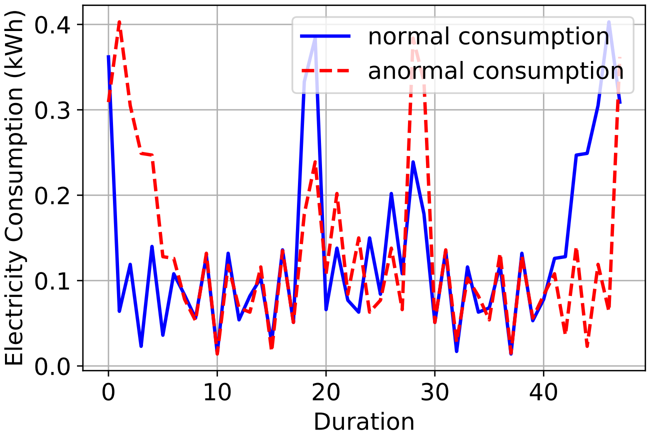

Figure 11: Fraud 10

11. **Fraud:** In this fraud, energy consumption is reduced only for a specified period, and the reduced amount is distributed to the remaining times of that day [45]. Thus, the overall consumption for this fraud will remain the same as the actual consumption, but the billing could be lower if the electricity company uses time-varying pricing systems. This type of fraud is only valid if there is no fixed pricing system. Eq. 4.11 and Figure 12 show the mathematical model for fraud 11 and its variation, respectively. Where, t_s is the starting time of the highest consumption n time period, N is total number of samples, $t_x = t_s + n$ and $\in = \sum_{j=1}^{n} x_{t_t+j-1}$.

$$f_{11}(t) = \begin{cases} x_t - \lambda x_t, t_s < t < t_x \\ x_t + \epsilon/N - n, \quad \text{else} \end{cases} \tag{4.11}$$

12. **Fraud:** This fraud occurs when an attacker switches their consumption pattern to a user with a low consumption pattern [45]. Thus, the legitimate user will unknowingly pay for the adversary's electricity. This type of fraud may also occur when attackers under-report their consumption and, at the same time, over-report the same proportion to their neighbors [46]. Frauds of this type cannot be easily suspected and detected. Figure 13 shows the fraud.

$$f_{12} = x_t = x_t * Z \tag{4.12}$$

Most frauds caused by injecting false data need attackers to have full or partial network information to modify the meter readings. Besides, attackers need to know

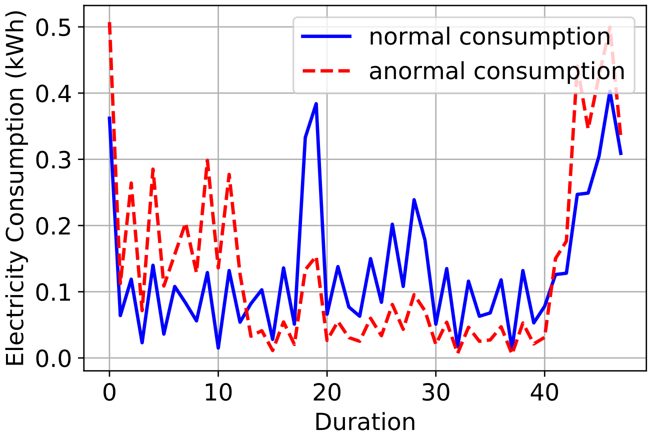

Figure 12: Fraud 11

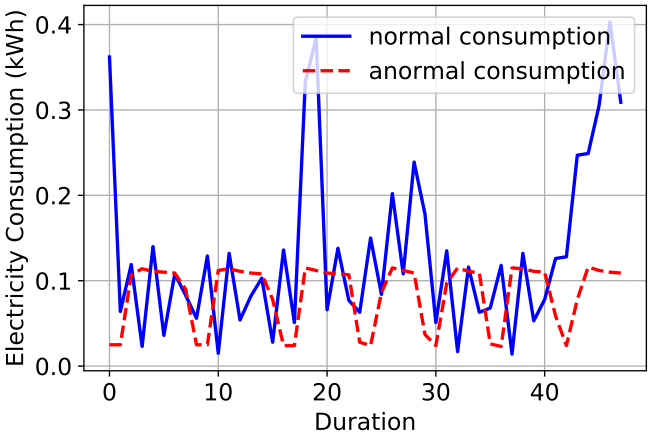

Figure 13: Fraud 12

the trend of real consumption before generating fraud vectors so that frauds can occur successfully with little chance of being detected by the electricity provider. In some frauds, attackers should be familiar with the pricing system used to maximize their benefit. Attackers often need help to gain full access to the SG network. Although there is a low probability of such frauds, they can not be ignored.

Once the fraud samples are generated, the data is labeled, consisting of n pairs of input and output values $(x1, y1), ..., (xn, yn)$. Conventional classifiers process time-series datasets as tabular data; each input value x pertains to a different feature, and features are selected according to their perceived relevance. Conventional ML classifiers can also utilize time series for resource extraction, a process in which a specialist analyses a time-series with filters and other specific techniques and extracts its most relevant predictors. On the other hand, TSC classifiers process datasets fully, consider their time dependency and decide which parts are appropriate or not. One of the advantages of using the time series is the removal of the human component from the process and the elimination of the resource extraction phase from the pre-processing step. Hence, the Algorithm considers the input values x, the entire daily time series, to predict output values y, i.e., the classes to which the series belongs.

4.1.2 Pre-Processing

Energy consumption data typically appears as a unidimensional time series with chronologically ordered readings. A critical step in the preprocessing stage involves adjusting this data. The algorithm utilizes a time series cross-validation technique to yield more reliable and robust results. Initially, it divides the data into training and test sets. Blocking time series then segments the data into n Folds, each designated with specific training and test sets. Every fold includes an identical number of samples, each divided sequentially. Subsequently, the algorithm trains classifiers with each training subset, selecting parameters that minimize classification error for each validation set. Finally, it configures the predictor with the optimal parameters and trains it using the entire training set. The aggregate performance is the average of the performances across all folds.

The *Blocking time series split* adds margins in two directions. One margin prevents the classifiers from memorizing future trends, and the other prevents the classifiers from re-memorizing patterns in between interactions. Despite the increased complexity of cross-validation, it is indispensable because it makes the predictors more error-resistant. This application needs to consider time series attributes when splitting data, guarantee that the predictors learn from past data (training stage), and make predictions about future data (test stage), making it better at working with real-world scenarios.

After the split, the process normalizes the training sets. Normalization adjusts all features to a uniform scale, ranging from 0 to 1. This normalization of entries prevents classifiers from favoring features with larger magnitudes, thereby enhancing the generalization capability of the predictor.

We *selected the parameters* of the ensemble learning grid search (i.e., tools for hyperparameter tuning. As mentioned, ML compares different models and tries to find the best one). Furthermore, they are chosen individually for each algorithm used in the ensemble. First, we maximize the efficiency of these algorithms individually. Then, we compare the impact of each algorithm on the ensemble by choosing the best set of parameters for them. Table 2 shows the parameters tested for each ML technique.

4.1.3 Ensemble Learning

Ensemble learning consists of methodologies that aim to deduce the specific function of the target by training various potential learners and integrating their hypotheses [18]. This study tested ensemble ML classifiers, incorporating boosting and bagging techniques, such as Catch22, Weasel, TSF, K-NN, and Arsenal.

It becomes necessary to evaluate the predictor when creating and training the Ensemble Predictor with the training set. Consequently, the trained predictor employs

Table 2: Table of Parameters

Classifier	Predictor

SVM	C = 100
	kernel = rbf
	gamma = 0.1

Signature	estimators = 100
	estimator = Random Forest
	window depth = 3
	depth = 4

CatBoost	estimators = 150
LightGBM	estimators = 150
XGBoost	estimators = 150

Catch	estimators = 200
	estimator = Random Forest

Arsenal	kernels = 1000
	transform = rocket
	estimators = 25
	estimador = Ridge Classifier

k-NN	neighbors = 1
	distance = DTW

Weasel	binning strategy = information gain window increment = 2 bigrams = true

the test set to forecast the labels of the test samples. Researchers then compare these predicted labels with the actual labels to assess the prediction performance. They present the results of this evaluation in a confusion matrix, forming the foundation for all evaluation metrics used to assess the Algorithm's effectiveness. The subsequent chapter will delve deeper into the analysis of these metrics.

4.2 Fraud Detection

4.2.1 Metrics

The F1-score is used to assess the data's positive predictive value and sensitivity to find some balance when using the harmonic mean. F1-score is the most appropriate metric for imbalanced datasets representing different class distributions among the five performance metrics used. The F1-score is the weighted average of the precision, where the first value is the ratio of the number of correctly predicted positive observations to the total number of predicted positive observations, as shown in Eq. (4.13).

$$F1 - Score = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{4.13}$$

Accuracy serves as a metric for evaluating classification models. Informally, it represents the proportion of predictions our model made correctly. Formally, accuracy is the ratio of correct predictions to total predictions. The false-positive rate is another metric for assessing the accuracy of machine learning models. A model must understand the "basic reality" or the actual state of affairs to measure its true accuracy. One can directly evaluate the model's accuracy by comparing its output against the ground truth.

$$Accuracy = \frac{TruePositives(TP)}{TruePositives + FalsePositives(FP)} \tag{4.14}$$

The False Positive Rate (FPR) assesses the incidence of false positives, where the system mistakenly classifies samples as fraudulent. FPR correlates with the additional costs a utility incurs when it erroneously dispatches teams for frequent inspection visits. Naturally, this situation also causes significant inconvenience for clients incorrectly identified as fraudulent.

$$FPR = \frac{FP}{FP + TrueNegative(TN)} \tag{4.15}$$

Recall measures the proportion of Actual Positives that our model correctly identifies as Positive (True Positive), according to Equation (4.16). With this understanding, Recall becomes the critical metric for selecting the best model when a high cost associates with a False Negative. For example, in fraud detection scenarios, failing to identify a fraudulent transaction (Actual Positive) as fraudulent (Predicted Negative) can lead to severe consequences.

$$Recall = \frac{TP}{TP + FalseNegative(FN)} \tag{4.16}$$

Precision metrics try to answer the question: how many attributes we identify correctly? Another way to express accuracy is the overall ratio of true and predicted positives. Precision considers the number of features correctly assigned to a given class versus the number of correct and incorrect assignments. Precision measures the classifier's correctness and the positive classification's correlation, which is computed based on Eq. (4.17). Higher precision means more true positives and fewer false positives.

$$Precision = \frac{TP}{TP + FP} \tag{4.17}$$

Area Under The Curve (AUC): AUC or ROC curves spans true positive and false positive rates and varies between 0 and 1, as in Eq. (4.18). ROC curves are an excellent metric for evaluating imbalanced databases. The higher the AUC, the more correctly the predictor can predict the outcome.

$$AUC = \int_{x=0}^{1} Recall(FPR^{-1}(x))dx \tag{4.18}$$

4.2.2 Results

Figure 14 shows the precision performance of all the classifiers implemented for NTL detection. By analyzing the results, we can conclude that the precision results closely follow the recall results, corroborating the possibility of adopting TSC classifiers for NTL detection. HybridForest shows better precision than the classifiers analyzed. It also provides precision results of around 80 %, which are 10 % and 14 % higher than CatBoost (*i.e.*, , the best-performing conventional classifier) and Catch (*i.e.*, , the best-performing TSC classifier), respectively. The precision performance metric shows that HybridForest is more likely to select a relevant sample randomly, *i.e.*, ., the number of hits returned that was TP or TN.

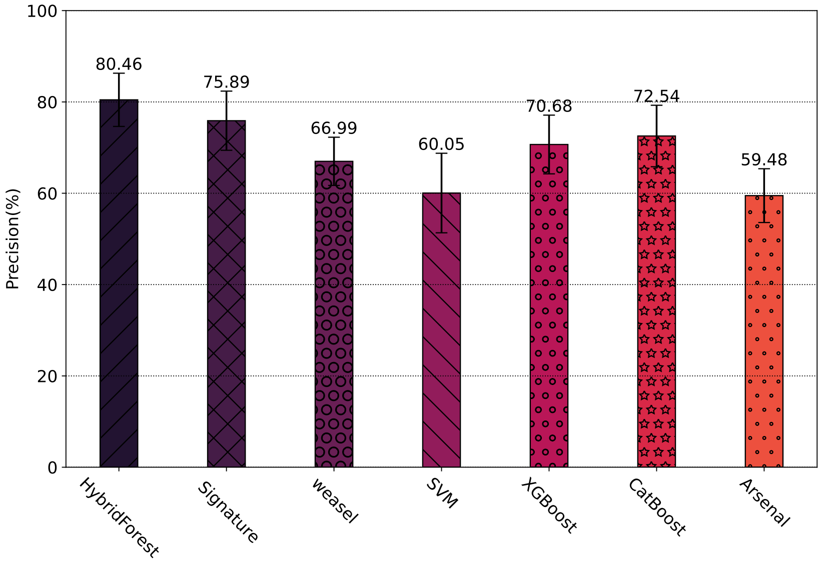

Figure 14: Precision Graph Bar

Figure 15 presents the F1-Score for the analyzed NTL detection predictor. Naturally, the F1-Score tends to behave similarly to the metrics used to calculate it. Accordingly, HybridForest provides a high ratio of correctly predicted positives to the total number of real positive samples, suggesting a high recall and precision. The F1-Score performance confirms the benefits of the HybridForest predictor compared to the predictors analyzed for NTL detection. Hence, the F1-Score results mean that HybridForest is efficient for both detecting frauds and for correctly identifying honest data samples because HybridForest considers the temporal nature of energy consumption data in the pre-processing, training, testing, and validation steps and also employs different TSC classifiers to create an ensemble predictor.

Figure 16 shows accuracy for the analyzed NTL detection predictor. Based on

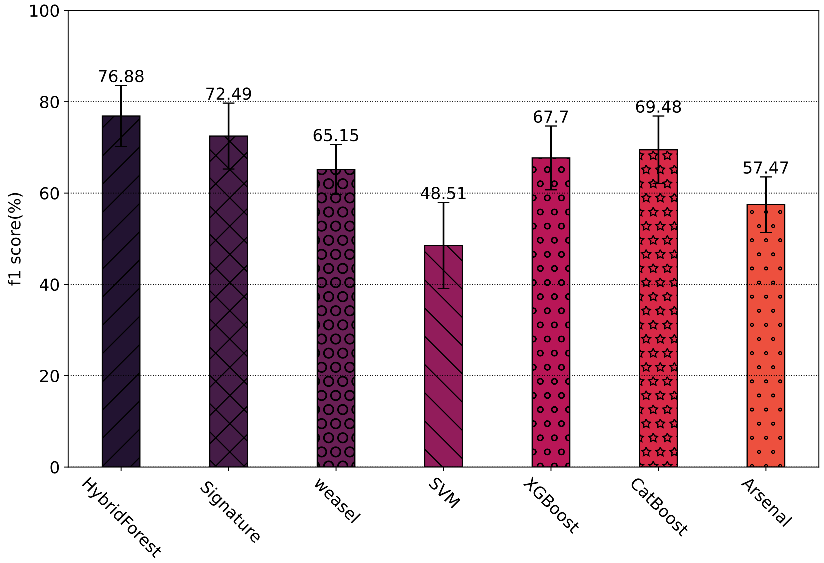

Figure 15: F1 Graph Bar

the results, we can conclude that the HybridForest has better accuracy than the classifiers analyzed. Specifically, HybridForest provides accuracy results of around 78 %, which is around 11 % and 10 % higher compared to CatBoost (*i.e.*, , the best-performing conventional classifier) and Signature (*i.e.*, , the best-performing TSC classifier), respectively.

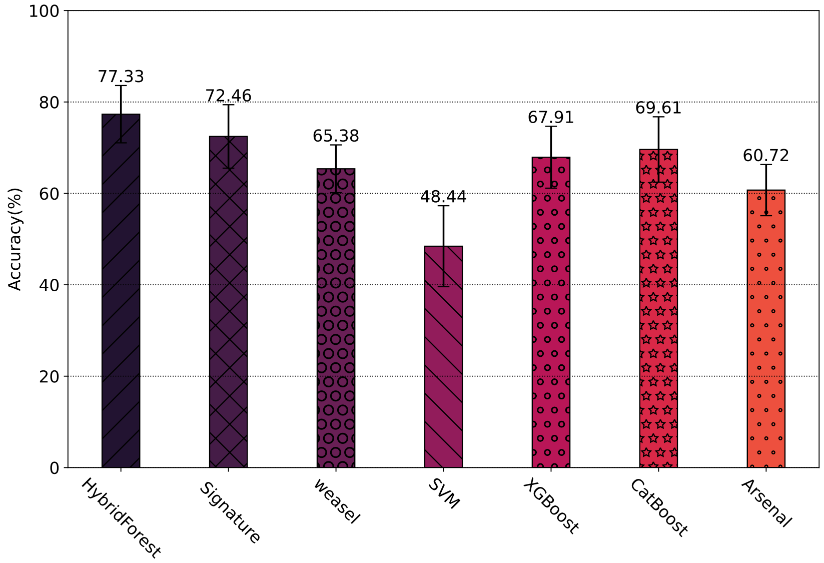

Figure 16: Accuracy Graph Bar

Figure 17 illustrates the FPR results, which directly correlate to the inspection costs incurred by utility companies. A small FPR value represents a small number of false detections. The FPR results also show the benefits of HybridForest for fraud detection compared to the predictors analyzed because HybridForest considers the temporal nature of energy consumption data and uses different TSC classifiers to create an ensemble predictor. For instance, HybridForest achieved a significantly lower FPR, 24 % and 35

% lower, compared to CatBoost (*i.e.*, , the best-performing conventional classifier) and Weasel (*i.e.*, , the best-performing TSC classifier), respectively. Those results corroborate our method's validity because it achieved very low FPR values compared to other NTL detectors, regardless of the classifier.

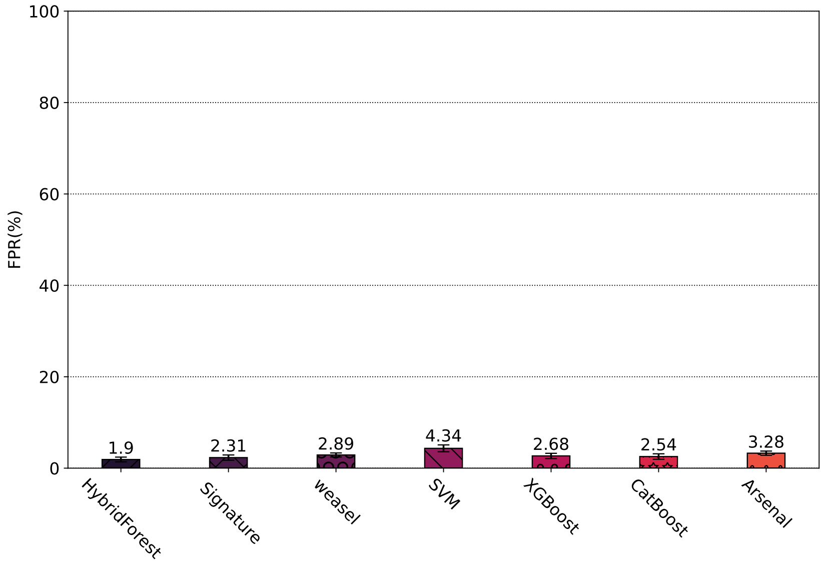

Figure 17: FPR Graph Bar

Figure 18 shows the recall performance of all the classifiers implemented for NTL detection. Recall is the most used performance metric for NTL detectors, and high recall values indicate that the predictor is efficient for fraud detection. By analyzing the results, we observed that all individual TSC-based predictors (i.e., Catch, Weasel, KNN, SVM, and Arsenal) performed worse when compared to conventional classifiers (i.e., XGBoost, CatBoost, and LightBoost), and HybridForest yields the highest recall performance. However, introducing an ensemble predictor (combining different TSC classifiers) produces a more robust and accurate predictor. Consequently, we can classify user samples with ease based on the recall.

Figures 19, 20 and 21 show the Easy ROC curves for each class predicted, and the Figures 22, 23 and 24 show the Hard ROC curves for each class predicted. The TSC-based predictors with the best performance (i.e., Signature) and the conventional classifiers (i.e., XGBoost). The ROC curve illustrates the relationship between recall and FPR, where an optimal result should achieve the highest possible recall value while maintaining the lowest possible FPR. We can observe that the performance varies among the different predictors and also among the different classes (i.e., honest data and twelve different types of fraud described by Eqs. 1-12). It is due to each class having different patterns and characteristics. Honest and fraudulent f1 data follow the same consumption pattern but with different amplitudes. It makes it difficult for predictors to distinguish one class from the other, especially when it started before the observation window [3]. The Fraud 4 has the worst performance regardless of the predictor. It is important to highlight that some predictors yield different performance results for frauds, e.g., HybridForest results

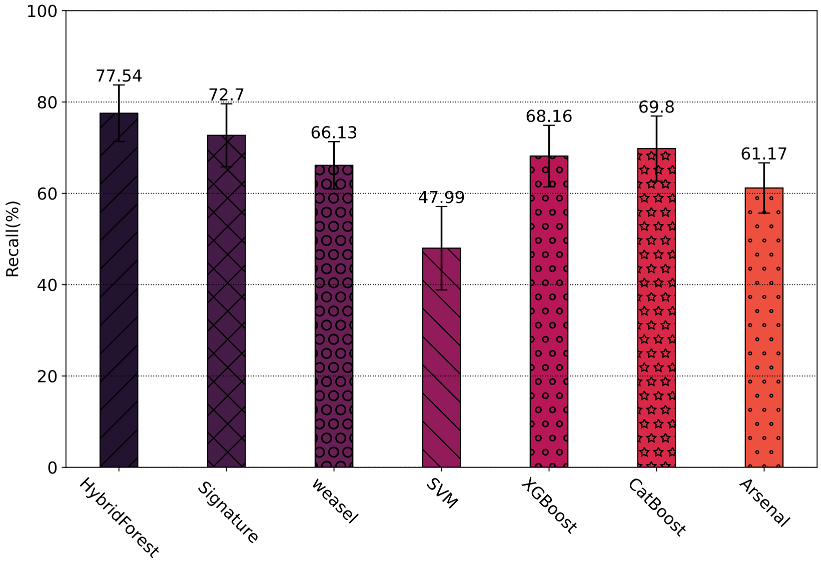

Figure 18: Recall Graph Bar

in an AUC of over 80% to identify Fraud 7, while XGBoost and Signature yield an AUC value of under 70% for the same class.

On the other hand, Fraud 9, a random cut-off point, is selected and subtracted from the actual sample [43], which facilitates prediction. In addition, the Fraud 12 can be easily detected because this fraud occurs when an attacker switches their consumption pattern to a user with a low consumption pattern [45]. In this way, for both Frauds 12 and 9, the Fraud patterns yield the best results, i.e., high recall and low FPR (close to 1), regardless of the predictor. Finally, Fraud 3 multiples each meter reading x_t by a different real pseudo-random number β_t between min and max values in a predetermined interval. Specifically, the fraud level increases as the βt value decreases. It means the NTL might not occur continuously, and there may be some discontinuous reporting of "fraudulent" values. Therefore, the user produces different reduction rates across measurements in this fraud pattern. It confuses conventional classifiers incapable of accounting for the time-dependent nature of the data and processing it point by point. Therefore, conventional classifiers will not be able to discern between this type of fraud and honest samples readily.

We can also observe that honest data and fraud patterns 1, 4, 6, 8, and 11 have the worst results compared to other classes (i.e., 2, 3, 5, 7, 9, 10, and 12), regardless of the predictors. For instance, the XGBoost predictor in Figures 21 and 24 show the worst results for honest data and fraud patterns in the 3 for easy and 4 for hard classes compared to the other ones. The curves for these classes take longer to reach the maximum value on the y-axis, which means that these classes have a lower recall than the others, i.e., a lower detection rate. On the other hand, detecting fraud patterns 2, 5, 7, and 9 for easy and 1, 6, and 8 for hard classes shows better performance, as their curves are higher than the average curve of all classes, be easier to make predictions for them, i.e., they have a higher detection rate.

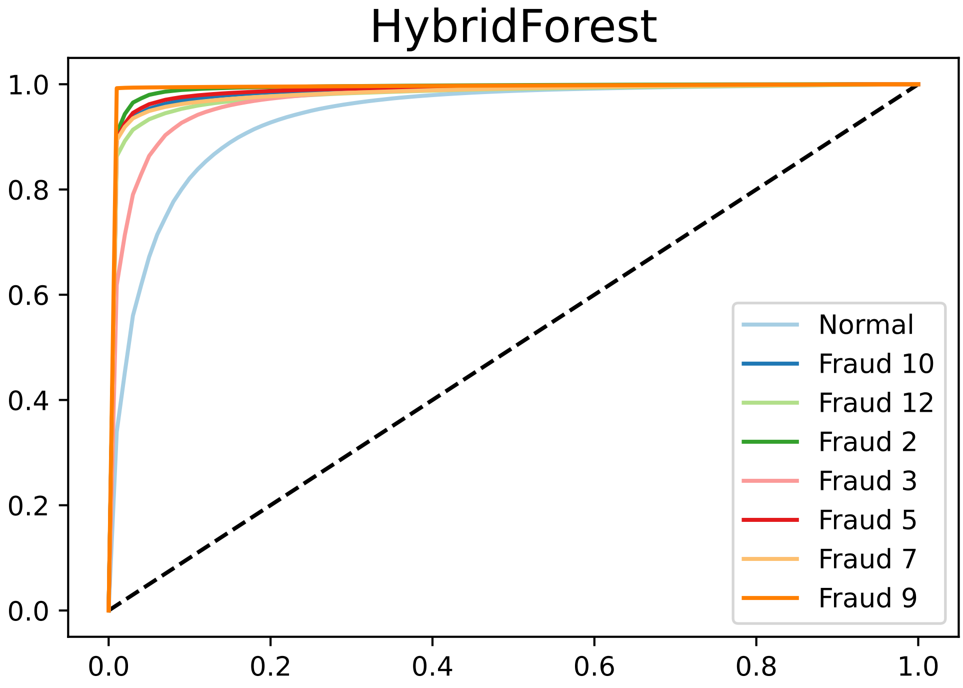

Figure 19: HybridForest Easy

By analyzing the results of each predictor, we conclude that HybridForest obtained high performances regardless of the classes, reaching the maximum y-axis value very quickly, especially for fraud patterns 2, 5, and 9 for Easy ROC and 1, 6, and 8 for Hard ROC. HybridForest relies entirely on time-series data processing combined with an ensemble predictor. On the other hand, the individual TSC-based predictor (i.e., Signature) also obtained acceptable results to detect each fraud pattern because they process time-series data. Lastly, the Conventional predictor (i.e., XGBoost) showed the worst performances because they treated the data in a tabular format, which can hinder classification for this application.

We conclude that HybridForest achieved a higher Recall, precision, and F1-score when compared to the other predictors. In addition, HybridForest outperformed all the other detectors in terms of FPR (i.e., 1.9), which directly translates into tangible benefits (such as the reduction of inspection costs) to utility companies as an expected trade-off for a more comprehensive and practical NTL predictor. Usually, binary NTL detectors group all fraud patterns in a single class, which can confuse the predictor by having different patterns as part of the same class, increasing FPR values. In this context, HybridForest took advantage of the time dependence inherent to time series in the classification process. Therefore, our methodology not only brings benefits to utility companies but also improves NTL detection. It is achieved by exploring the available data in novel ways and by employing and combining the most recent resources for classification problems more efficiently to detect specific types of NTL.

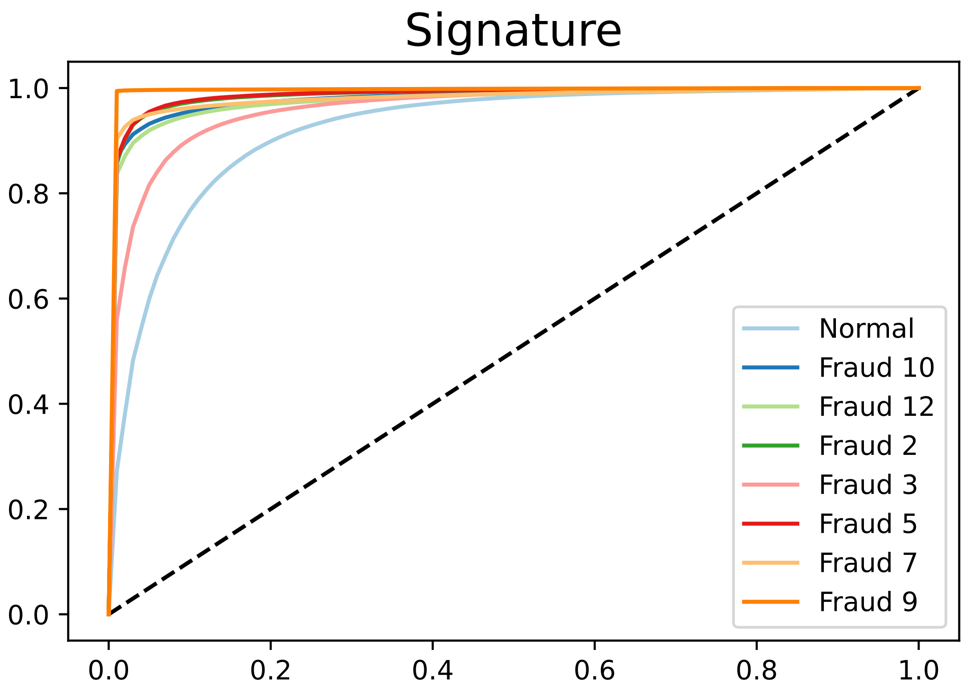

Figure 20: Signature Easy

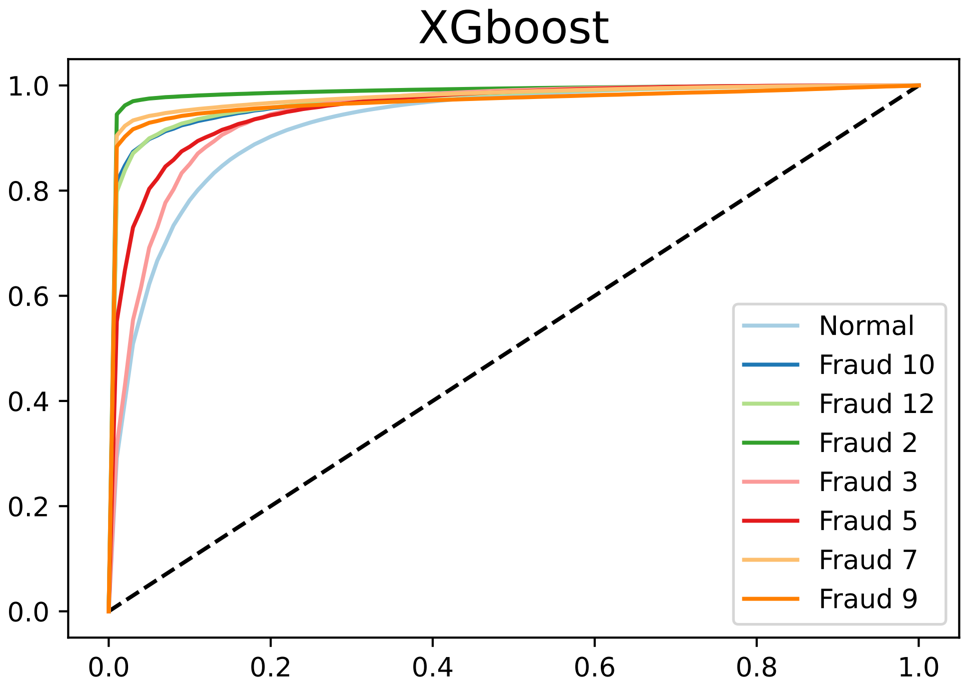

Figure 21: XGBoost Easy

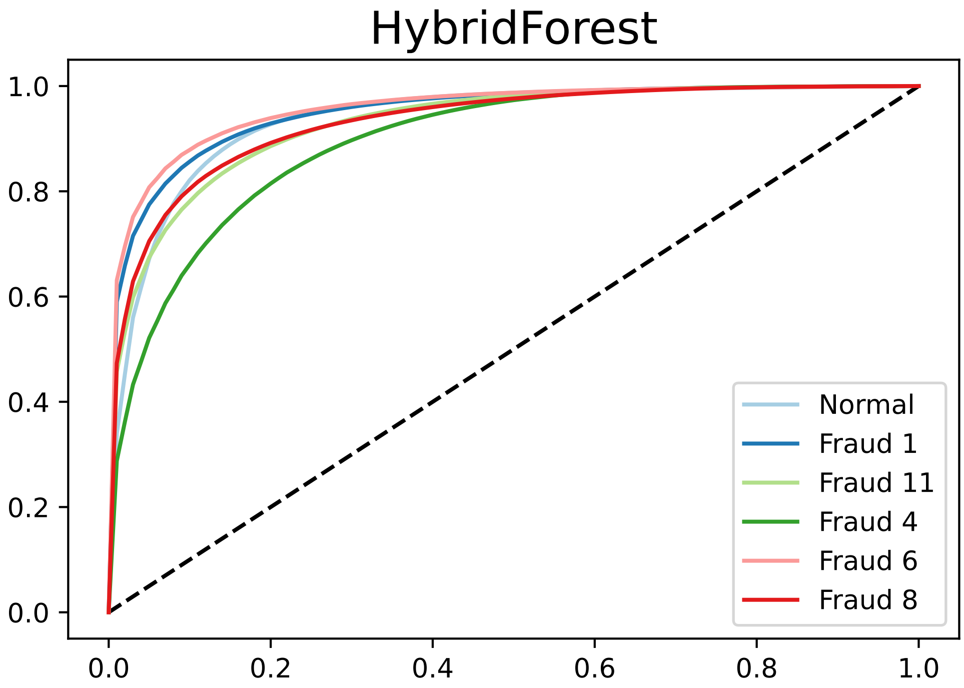

Figure 22: HybridForest Hard

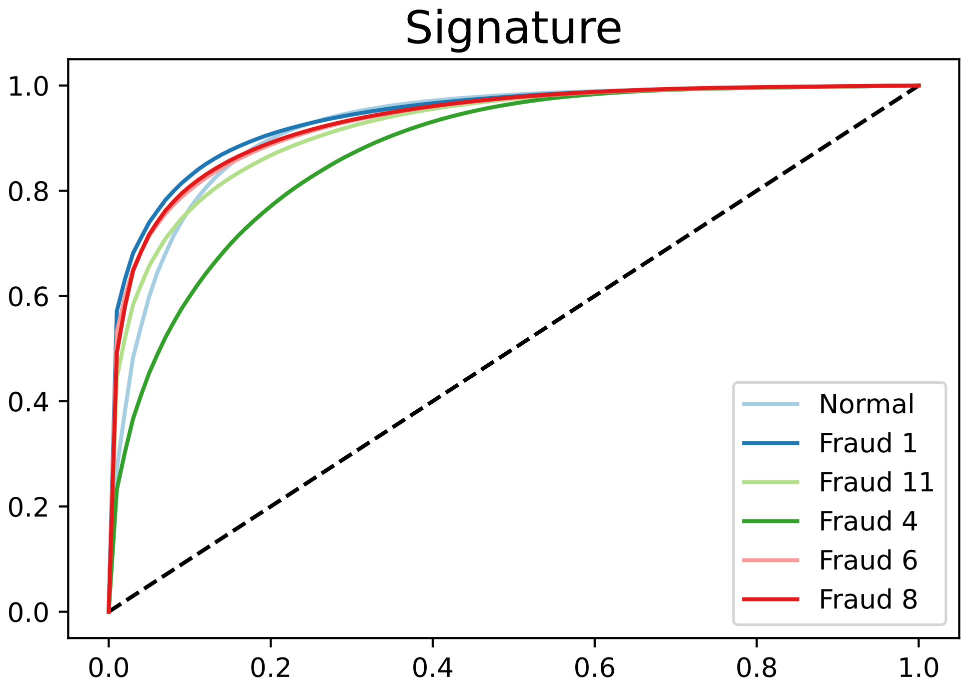

Figure 23: Signature Hard

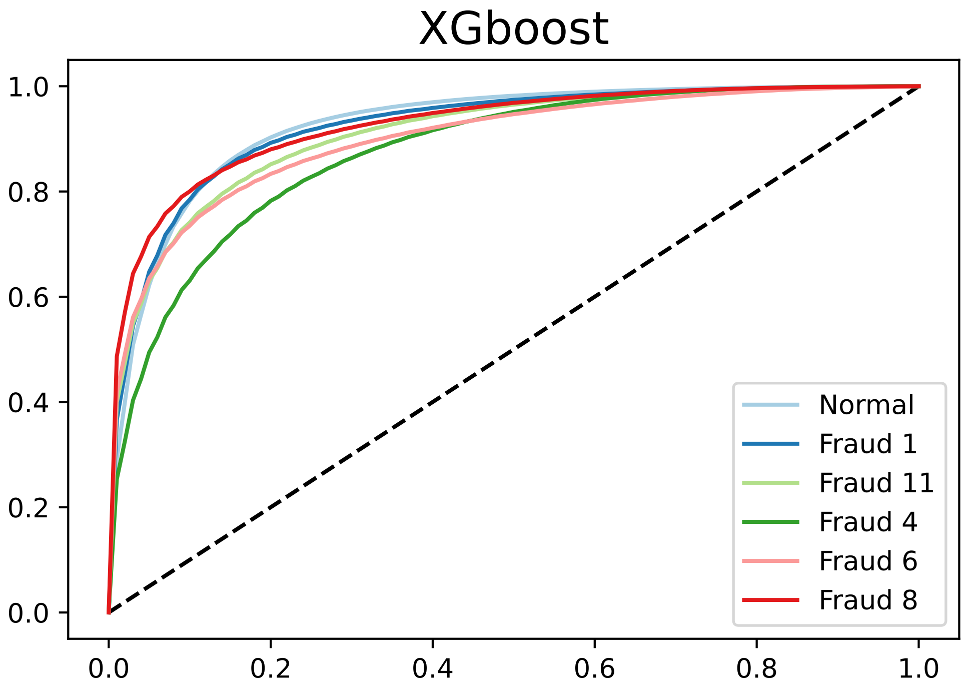

Figure 24: XGBoost Hard

4.3 Chapter Conclusions

Despite many efforts to detect fraud, it is still an open issue. In this thesis, we presented a data-oriented heterogeneous ensemble predictor for NTL detection in an SG scenario. HybridForest relies on the heterogeneous Ensemble to perform a multi-classification of fraudulent users (i.e., classifying samples as honest or as a specific type of fraud) with high performance against other methods such as SVM, XGBoost, CatBoost, and LightBoost to identify types of fraud. In addition, we consider that these frauds in the electrical system are not only binary, and our predictor can also classify these variations among them and their different unique aspects. In this context, HybridForest considers the temporal nature of energy consumption data in the pre-processing, training, testing, and validation stages and different TSC classifiers to create an ensemble predictor.

For evaluation, we use the Irish dataset, which has a large number and type of users and a long duration of measurements. This database includes only honest consumption data because of its scenario of uninterrupted monitoring of known users who earlier accepted the terms of the compact and who had to respond to a questionnaire before and after the measuring period. We counted twelve types of fraud already described in the literature in this dataset to create a synthetic dataset with honest and fraudulent consumers. In this context, we can determine different types of fraud, whereas existing related works only make binary categories (fraudulent and non-fraudulent). We also followed the method adopted in these works, with fraudulent data randomly selected among users. It is essential because fraudulent samples need to be more balanced and need to be used by ML algorithms.

This thesis has presented a data-oriented ensemble predictor for NTL detection in an SG scenario. The algorithm employs a time series of user consumption data to build a predictor to classify samples as honest or a specific type of fraud. We tested and compared multiple TSC algorithms in our experiments. The TSC algorithms performed better than the conventional classifiers for all metrics, demonstrating the benefits of using this classifier to create a prediction for NTL detection. By employing TSC classifiers to build an ensemble predictor HybridForest, we obtained a performance improvement with an FPR value equal to 1.9% and a precision 80.5% for heterogeneous data and kinds of frauds. The algorithm focuses on time-series data, which enables the development of a method that better interprets real-world scenarios and is more error-resistant.

CHAPTER 5

Performance Evaluation of Frauds Characterization based on Information Theory Quantifiers

This chapter presents the results of data-based characterization frauds based on information-theoretic measures. We attempt to characterize different kinds of fraudulent users in energy consumption. We describe the evaluation by first explaining the methodology used in the simulation and its parameters and metrics. Then, we discuss the obtained results.

5.1 Evaluation of Characterization

This section discusses the main ITQ used in determining a time series's chaotic and stochastic nature, such as the speed dynamics of a means of transportation. We also explain BP's symbolization method since evaluating the ITQ in such time series requires defining probability distributions associated with them. Finally, we detail the causal planes: CECP and Fisher-Shannon Causality Plane.

We used the BP method [13] to transform raw electricity consumption into a histogram. Specifically, the BP symbolization method assigns probability distributions from the time series under consideration, *i.e.*, , the temporal causality of the process. In this sense, given a time series $\boldsymbol{X}(t) = \{x_t : t = 1, \dots, N\}$ (*i.e.*, , energy consumption data), an embedding dimension $D \geq 2 (D \in \mathbb{N})$, and an embedding delay time $\tau \in \mathbb{N}$, we compute the ordinal patterns of order D (pattern length) generated by

$$(s) \mapsto \big(x_{s-(D-1)\tau}, x_{s-(D-2)\tau}, \ldots, x_{s-\tau}, x_s\big) \tag{5.1}$$

Afterward, we assign each point in time s with a D-dimensional vector resulting from evaluating the sequence at time $s-(D-1)\tau$, $\ldots, s-\tau, s$. More information about the past is built into the vector by considering a higher D value. According to the pattern, the meaning of order D with respect to time s is permutation $\pi = \{r_0, r_1, \ldots, r_{D-1}\}$ of $\{0, 1, \ldots, D-1\}$ is defined as

$$x_{5-r_{D-1}\tau} \leq x_{5-r_{D-2}\tau} \leq \cdots \leq x_{5-r_1\tau} \leq x_{5-r_0\tau} \tag{5.2}$$

Thus, the data produced by Eq. 5.1 is converted to the unique symbol π. To get unambiguous results, we set $r_i < r_{i-1}$ if $\chi_{s-r_i} = \chi_s r_{i-1}$. If $X(t)$ follows a slightly continuous distribution, then the probability of equal values is zero. Therefore, we calculate the associated relative frequencies for all $D!$ and the possible permutations π of order D, which this particular ordered sequence was found in the time series divided by the total number of sequences. Hence, the histogram $P \equiv \{p(\pi)\}$ is defined as

$$p(\pi) = \frac{\#\{s \text{ of type } \pi : s \leq N-(D-1)\tau\}}{N-(D-1)\tau}, \tag{5.3}$$

where # is the cardinality of the set.

The second concept is Shannon Entropy, a global measure of self-information. Let $\mathcal{X} = \{x_j : j = 1, \ldots, M\}$ be a discrete random variable of length $M < \infty$ whose distribution features is the probability function $P = \{p_i : i = 1, \ldots, M\}$. p_i represents the probability of state i, and $\sum_{i=1}^{M} p_i = 1$, and M is the number of possible states of the checked system. The well-known Shannon entropy is

$$S[P] = -\sum_{i=1}^{M} p_i \ln p_i, \tag{5.4}$$

Among them, $p_i \ln p_i = 0$ if $p_i = 0$, it is related to the physical process described by P. Once the Shannon entropy $S[P] = 0$, the information (knowledge) of the underlying process described by P is maximal and possible outcomes can be predicted with complete certainty. On the other hand, if the physical process follows a uniform probability distribution $P_e = \{p_i = 1/M, \forall i = 1, \ldots, M\}$ then little knowledge is obtained [13].

It is also helpful to define the so-called normalized Shannon entropy to evaluate the self-information in a normalized way, denoted by

$$\mathcal{H}[P] = \frac{S[P]}{S_{max}} = \frac{S[P]}{S[P_e]} = \frac{S[P]}{ln\, M}. \tag{5.5}$$

We need to find a proper measure of complexity based on classification, or information alone. In this case, Lamberti *et al.* [47] proposes an SC measure $\mathcal{C}_{JS}[P]$, which can identify important dynamic details, such as electricity consumption, a profile of users based on consumption, and others. López-Ruiz *et al.* [48] proposed this complexity measure based on the product of functions,

$$\mathcal{C}_{JS}[P] = \mathcal{H}[P]\, \mathcal{Q}_{JS}[P, P_e], \tag{5.6}$$

Where $\mathcal{H}[P] \in [0, 1]$ is the normalized Shannon Entropy, and $\mathcal{Q}_{JS}$ is the disequilibrium based on the Jensen-Shannon (JS) divergence. In this sense, $\mathcal{Q}_{JS}$ is expressed by

$$\begin{aligned} \mathcal{Q}_{JS} &= Q_0 \mathcal{J}_S[P, P_e] \\ &= Q_0 \left\{ S\left[\frac{P + P_e}{2}\right] - \left[\frac{S[P] + S[P_e]}{2}\right] \right\}, \end{aligned} \tag{5.7}$$

Where Q_0 is a normalizing constant, while $\mathcal{J}_S$ is the JS divergence to quantify the difference between probability distributions. The presence of correlation structure is quantified in the SC [49], which measures time series complexity. In the case where the signal from the dynamical system is ultimately ordered or completely random, the value of $\mathcal{C}_{JS}[P]$ is the same as `null`, *i.e.*, , the signal has no structure. In between these two extremes, dynamic systems can perform every possible level of physical structure. These phases should be reflected in the obtained features of the Probability Density Function (PDF) and quantified by `no-null` $\mathcal{C}_{JS}[P]$. The global property of SC is that its value does not change with different PDF layouts. Thus, $\mathcal{C}_{JS}[P]$ quantifies disorder but also the degree of correlated structure.

The third concept used in our characterization is the *FI*, which is used to analyze local aspects of changes in the information content given by a time series. It has different interpretations and calculations; among other things, the amount of information extracted from a process is a measure of the ability to estimate parameters or the disordered state of a system or phenomenon [49]. We define it as

$$FI[P] = F_0 \sum_{i=1}^{N-1} \left(\sqrt[2]{p_{\mathrm{i+1}}} - \sqrt[2]{p_i}\right)^2, \tag{5.8}$$

Where F_0 is a normalization constant defined by

$$F_0 = \begin{cases} 1 & \text{if } p_{i^*} = 1 \text{ for } i^* = 1 \text{ or } i^* = M \\ & \text{ and } p_i = 0\, \forall i \neq i^* \\ 1/2 & \text{otherwise.} \end{cases} \tag{5.9}$$

According to Olivares *et al.* [49], the local sensitivity of FI to discrete PDFs requires the order of i of discrete values in $P = \{p_i : i = 1, \ldots, N\}$ when summing from Eq. 5.8. It is a distance between two related probabilities. Therefore, different orders will result in different FI values and, thus, their local nature.

Finally, we use the concept of the Causal Informational Plane in our characterization to allow a visual interpretation. The result of the BP symbolic approach is a probability distribution (P), also known as PDF-BP or BP-PDF, based on ordinal patterns described by time series. In this sense, according to Olivares *et al.* [49], we can use two representation spaces to characterize a given dynamical system described by a time series: (a) One has a global-global property called CECP, and (b) has global-local features called Fisher-Shannon Causal Plane (FSCP) or FS plane.

CECP is a two-dimensional representation obtained by plotting a given system's permutation SC (vertical axis) versus permutation entropy (horizontal axis). The permutation SC refers to the complexity measure ($\mathcal{C}_{JS}$) proposed by Lamberti *et al.* [47] on the distribution P. Similarly, the normalized PE is the normalized Shannon entropy applied to the BP-PDF.

This level is particularly effective in distinguishing the deterministic chaotic and stochastic nature of time series since permutation quantifiers have specific behavior for different dynamics [13]. For each entropy value, we bound the plane by the minimum and maximum complexity curves ($\mathcal{C}_{JS}^{min}$ and $\mathcal{C}_{JS}^{max}$).

The Fisher-Shannon causal plane is a two-dimensional plot obtained by plotting the *FI* for a given system against the normalized permutation entropy (horizontal axis). The term causality is due to the temporal correlation between consecutive samples. In this case, according to Olivares *et al.* [49], in a system with M distinct states reaching a very ordered state, we can think of it as producing a time series whose PDF by $P_0 = \{p_k \cong 1,$ and $p_i \cong 0; \forall k \neq i = 1, \ldots, M\}$, because there exists Shannon entropy $S\left[P_0\right] \cong 0$ and normalized $FI\left[P_0\right] \cong F_0 = 1$. On the other hand, if the system to be analyzed develops into a very disordered state, it is reasonable to assume that a PDF describes this particular state that approximates the uniform distribution $P_e = \{p_i = 1/M; \forall i = 1, \ldots, M\}$, and corresponding Shannon entropy $S\left[P_e\right] \cong S_{\max} = \ln\, M$ while $F\left[P_0\right] \cong 0$.

As a result, we considered these ITQ techniques to characterize the frauds in the informational planes, as shown in the following workflow (as shown in Figure 25):

1. We use a sliding window to convert a user's energy consumption time series to the BP probability distribution function in the following way: we map to a unique symbol π in the histogram the pattern of a certain amount of samples, given by D, within a window specifies. Then, a probability mass function $p(\pi)$ is associated with each symbol according to the number of occurrences of a specific pattern in the entire series.

2. We extract the *FI*, PE, and SC ITQs used by the technique.

3. We apply each metric to the obtained BP PDF and later map it to the CECP and Fisher-Shannon planes characterizing the frauds.

4. Finally, each step previously described is repeated for each user in the dataset.

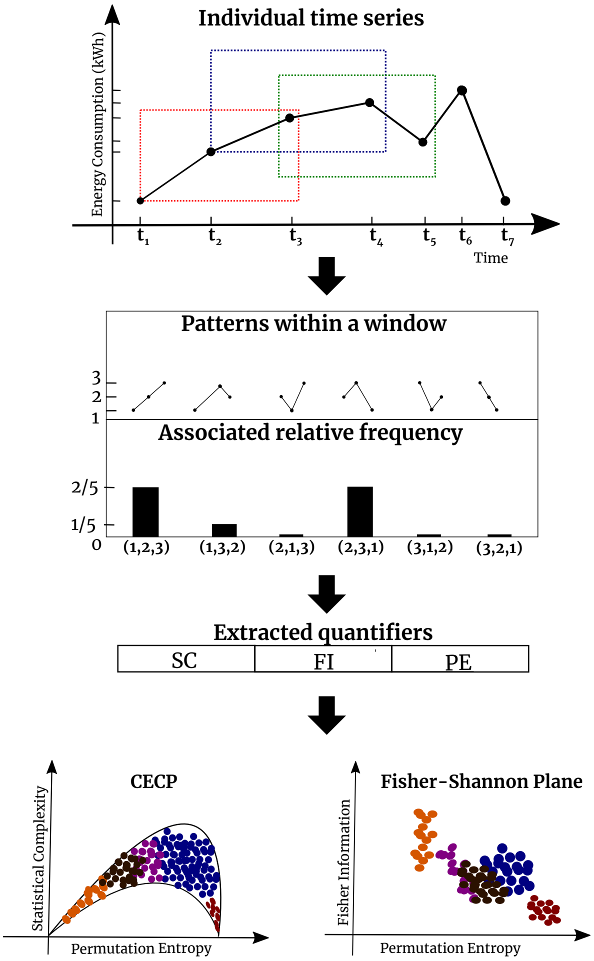

Figure 25: Characterization of Frauds with ITQ

It is difficult to find a public dataset with labels about honest and fraudulent users for data classification, and to the best of our knowledge, there is no public dataset with such information. In this context, we use consumption data obtained in a constantly monitored scenario wherein the participants accepted the terms of commitment to have an SM installed in their homes. It is reasonable to assume that all samples belong to honest users [41]. We used the dataset from the Ireland Smart Metering Energy Project [12] for the implementation. The dataset includes more than 5000 private and commercial electricity consumers from 2009 to 2010. The SM reports 48 energy uses in a 30-minute interval report daily. During the pre-processing, daily reports with less than 48 measures were removed since keeping them might result in a more significant number of outliers in the CECP and FS plane.

Similar to other works that use the Irish dataset, we add five fraud types already defined in the literature to create a synthetic dataset of honest and fraudulent customers. In this way, we generate synthetic fraudulent consumption data for the whole time series, from the first to the last reading, because we considered consumption data composed of energy consumption samples belonging to honest users, as other works [30, 34, 1] have also done. We generate samples of five fraudulent consumption patterns based on Jokar et al. [11], where malicious samples are generated by considering a continuous or periodic decrease in energy consumption.

In such a synthetic dataset, 40% of users were chosen randomly to change their normal behavior for fraudulent behavior. Specifically, in each fraudulent user, we replaced 25% of its energy consumption with an abnormal one given by the mathematical model of the respective fraud. The amount of samples for each fraud in the dataset is smaller since those classes would be less frequent than the normal class on a real problem. In general, frauds report less energy than the actual energy consumed or redistribute the energy consumption at different times to take advantage of the varying billing system [1]. Specifically, we consider five models of synthetic NTL attacks proposed by recent works [3].

Fraud 1 ($\boldsymbol{m_1}$) occurs when we multiply the actual consumption (e_t) by various random factors (γ_t) between 0 and 1. Each instance has a unique trend and size where the aggression level increases as γ_t decreases. We define it as

$$m_1(t) = \gamma_t * e_t, \tag{5.10}$$

Fraud 2 ($\boldsymbol{m_2}$) occurs when the attacker does not want to exceed the maximum predefined limit. We define it as

$$m_2(t) = \begin{cases} e_t, & e_t a \\ a, & e_t > a, \end{cases} \tag{5.11}$$

where a is cutoff point and $a < \max(e_t)$. The attacker chooses a random cutoff point, where all energy consumption values above this are reduced because it is fraudulently changed to the cutoff value [3]. The attacker chooses the cutoff point carefully, as a low cutoff point will cause the meter to report a series of easily identifiable constant consumption. On the other hand, if the threshold is too high, the amount stolen will be too low to be of any use to the attacker.

Fraud 3 ($\boldsymbol{m_3}$) occurs when a random cutoff point is selected and subtracted from the sample. We define it as

$$m_3(t) = \max(e_t - a, 0) \tag{5.12}$$

If the result obtained is less than zero, it reports zero. This attack could be due to incorrect data entry or bypassing the meter by connecting the load directly to the distribution transformer. Therefore, the meter does not record the amount consumed by the load.

Fraud 4 ($\boldsymbol{m_4}$) is define as

$$m_4(t) = \text{mean}(e_t) \tag{5.13}$$

It is easily detected by analyzing consumption profiles since the specification of average power consumption (constant power consumption) does not reflect the average power consumption of customers; it must vary randomly over time. This attack has no trend, mainly due to wrong data injection.

Fraud 5 (m_5) occurs when the energy consumption is reduced only for a specific time, and it distributes the amount of reduction over the rest of the day. We define it as

$$m_5(t) = \begin{cases} e_t - \lambda\, e_t, & t_s < t < t_e \\ e_t + \epsilon/N - n, & \text{else} \end{cases} \tag{5.14}$$

where the specific time occurs between the time, start (t_s) and time end (t_e), λ is the reduction rate, N is the total number of samples, $t_e = t_s + n$ and $\epsilon = \sum_{j=1}^{n} e_{t_s+j-1}$. Therefore, the total consumption for this attack is the same as the actual consumption, but the billing may be lower if the utility uses a time-varying pricing system. This type of attack is only effective without a fixed price system.

We implemented and analyzed our fraud characterization results using quantifiers from ITQ in Python software. We used ordpy, a pure Python module that implements data analysis methods based on the BP symbolic encoding schemes. We used the following characterization. First, the raw electricity consumption time series is transformed into a histogram containing causal time information using a non-parametric transformation: the BP method, for which we used $D = 6$ and $\tau = 1$. Second, we map this histogram to the CECP, and its position represents many canonical states. The plane is a compact manifold containing values of normalized Shannon entropy $\mathcal{H}$ and SC $\mathcal{C}$. Later, we observe the distinct frauds detected in a Fisher-Shannon Plane.

5.2 Results

Figure 26 exhibits the heatmap of the normalized permutation entropy for all the normal consumption and the five distinct frauds (*i.e.*, , m_1, m_2, m_3, m_4, and m_5) when considering each user's electric consumption behavior. Normal consumption and m_1 fraud aggregate near the highest values for nPE (*i.e.*, , in the interval 0.85 and 1 nPE), while almost all m_4 detected fraud is reunited near the range 0 and 0.05, thus presenting a minimum entropy and maximum certainty. The other three frauds appear around distinct normalized permutation values, *e.g.*, , over 83% of the detected m_2 frauds are around the interval 0.75 and 0.90; over 82% of m_3 frauds share a similar nPE result around the 0.40 and 0.55 interval; and over 87% m_5 frauds are close to interval 0.60 and 0.70. These results attest that m_1 and normal consumption share some similarities and might indicate that such fraud might be harder to detect apart from the normal consumption due to both profiles presenting maximum entropy. However, when analyzing others' NTL behaviors, the other frauds are much more spread when analyzing their entropies, and thus, they are easier to detect.

Figure 27 shows the heatmap of a FI distribution among all the detected users. In such evaluation, all the user's consumption presented values aggregated in a closer range (*i.e.*, , lesser than 0.50 FI). Normal and m_1 users showed similar results toward smaller FI values (*i.e.*, , the majority of values located around 0-0.15 FI), while m_4 detected

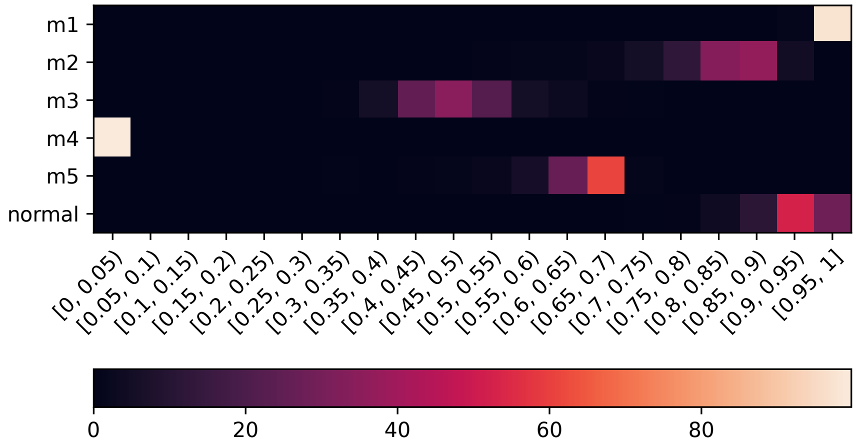

Figure 26: nPE values distribution (%)

users are highly concentrated around the range 0.35 and 0.40. The other three frauds, m_2, m_3, and m_5, are more spread toward middle FI range values (*i.e.*, , 0.15-0.30). These results also indicate that normal consumption and m_1 share similar FI values, suggesting that their electric consumption presents local similarities regarding information aspects. Hence, other NTL frauds are harder to detect when only analyzing FI results, demanding further evaluations.

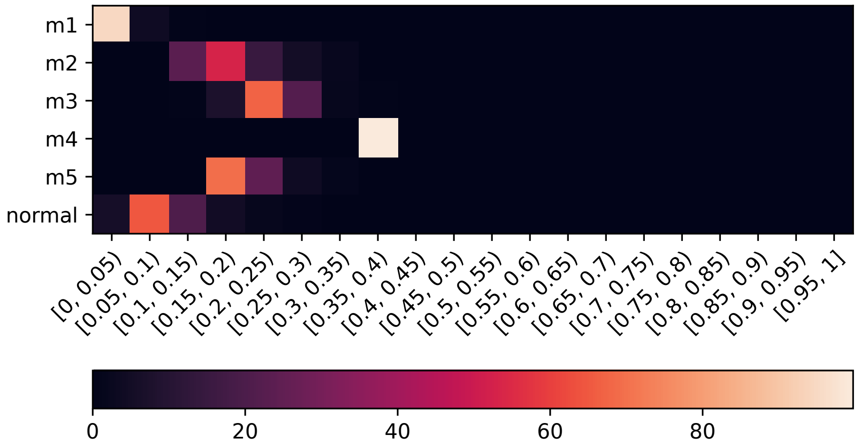

Figure 27: FI values distribution (%)

Figure 28 presents the heatmap of the SC results for all the intervals when analyzing the normal and NTL electric consumption behavior. The majority of values are around lesser values of SC (*i.e.*, , from 0-0.40). By analyzing all the normal and the five frauds' electric consumption behavior, the SCs concentrate on similar results, which indicate that it is harder to differentiate between the probability distributions. Therefore, further discussion is necessary when analyzing CECP and FS plane metrics.

Figure 29 evaluates all the distinct electric consumption time series from the distinct users. As stated before, we opted to detect five distinct frauds among the normal electric consumption of the users. By analyzing the CECP results, we characterize each electric time series according to its SC versus permutation entropy behavior. Depending on the behavior detected, some users' consumption might indicate a normal consumption,

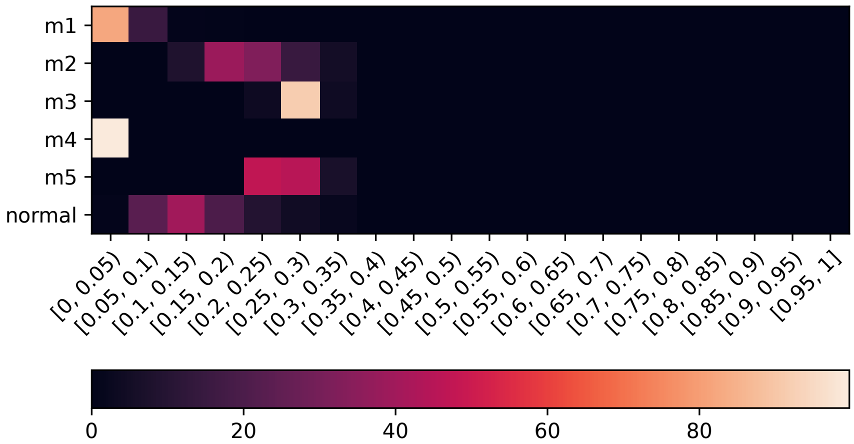

Figure 28: SC values distribution (%)

i.e., , an expected consumption. However, we identified other time series as distinct frauds, *i.e.*, , m_1, m_2, m_3, m_4, m_5. These results state that m_1 values are near the highest nPE and lowest SC values. At the same time, normal consumption is spread around the SC results but concentrated with maximum entropies. The m_4 fraud appears near the smallest values on the axis, while other frauds are spread but easily detected due to CECP. Therefore, the CECP presented good results when evaluating more than one ITQ.

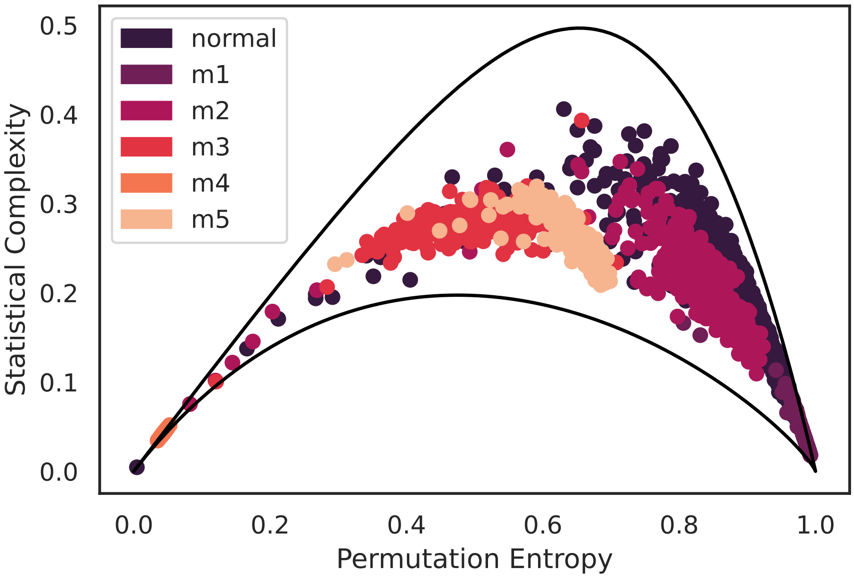

Figure 29: CECP evaluation.

Figure 30 compares all the time series from the dataset according to a Fisher-Shannon plane behavior. It enables the evaluation of the growth regarding the PE x FI results in a two-dimensional diagram and considers the temporal correlation between successive samples. It is clear to note the grouping of each user according to its normal consumption or among the five distinct detected frauds used, *i.e.*, , m_1, m_2, m_3, m_4, and m_5. These results state that each profile can be detected when evaluating its

Fisher-Shannon plane since it uses global-local characteristics from their position in the plane. Therefore, when analyzing all the regular and fraudulent users, we can see the patterns of similarity when using the ITQ and when uniting the analysis of the planes. Such characterizing distinct users by observing their electric consumption indicates the proposal's effectiveness in detecting NTL.

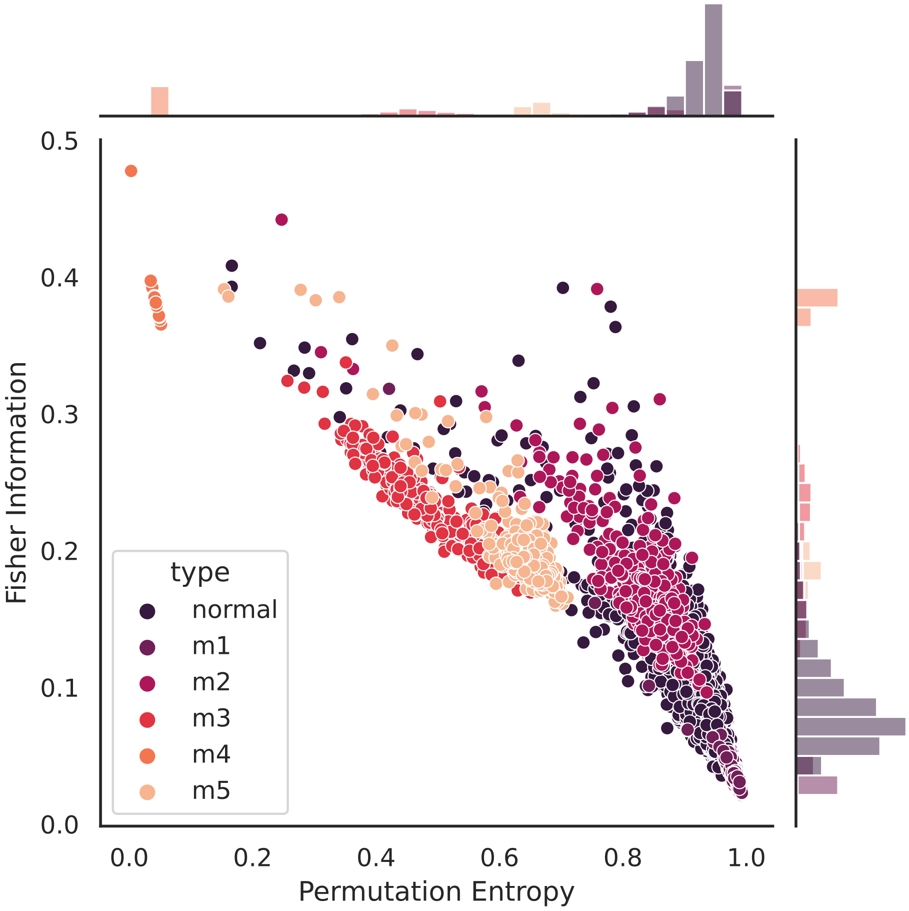

Figure 30: FSCP.

Using only ITQ, our results show that we can characterize users with normal and abnormal energy consumption by considering the values for each metric and also characterize types of frauds since users with similar behavior are around the same interval of the proposed quantifiers. As we can see in the plan, fraudulent users group with their respective types of fraud. Thus, these users vary their Permutation Entropy and Fisher Information similarly.

5.3 Chapter Conclusions

We introduced a correctly fraudulent characterization of users associating ordinal patterns, ITQ, stochastic processes, and causal information planes. Thus, we can precisely classify each fraud and analyze the user. However, we can find patterns of similarity between these frauds to characterize them better and associate them with compatibility between them, which might prove helpful in the future modeling of random patterns for different datasets. We can also test our method with other scenarios and types of fraud.

CHAPTER 6

Conclusion

This thesis proposal introduces two fraud characterization models for the electric power system, highlighting the effectiveness of ensemble learning in detecting various types of fraud and achieving superior metric results compared to other studies. And the other hand, the application of information theory techniques facilitates more fluid data handling, minimizes processing time, and maximizes data volume to enhance metric outcomes.

Despite numerous attempts to combat fraud, it remains a persistent challenge. This thesis proposes a data-oriented, heterogeneous ensemble predictor, HybridForest, for Non-Technical Loss (NTL) detection within a Smart Grid (SG) scenario. HybridForest employs a heterogeneous ensemble to conduct multi-classification of fraudulent users, distinguishing between honest and specific types of fraud. This approach demonstrates high performance relative to other methods, including SVM, XGBoost, CatBoost, and LightBoost, in identifying fraud types. Furthermore, HybridForest acknowledges the complexity of fraud beyond binary classifications, enabling the differentiation of various fraud nuances and their distinctive features. It incorporates the temporal dynamics of energy consumption data throughout the preprocessing, training, testing, and validation phases, employing diverse Time Series Classification (TSC) classifiers to establish an ensemble predictor.

The evaluation employed the Irish dataset, characterized by its extensive user base, lengthy measurement duration, and exclusive inclusion of honest consumption data, thanks to continuous monitoring and participant consent. To simulate a mixed dataset of honest and fraudulent consumers, we introduced twelve fraud types identified in literature into this dataset. This approach allows for recognizing diverse fraud types, moving beyond the binary classifications prevalent in related research. Including randomly selected fraudulent data among users is crucial to maintaining sample balance and suitability for

Machine Learning (ML) algorithms.

This thesis has developed a data ensemble predictor for NTL detection in an SG scenarios, leveraging consumption data time series to distinguish honest from fraudulent consumption patterns. Through experimental comparison of multiple TSC algorithms, TSC classifiers outperformed traditional classifiers across all metrics, underscoring their advantage in predicting NTL. Applying TSC classifiers in creating the HybridForest ensemble predictor enhanced performance, with an FPR of 1.9% and a precision of 80.5% across heterogeneous fraud types and data. This focus on time-series data paves the way for a methodology that more accurately reflects real-world conditions and exhibits greater resilience to errors.

We also introduced an accurate fraud characterization method for users, combining ordinal patterns, ITQ, stochastic processes, and causal information planes. It allows for precise fraud classification and user analysis while identifying similarities between frauds for better characterization and potential future compatibility modeling. The versatility of our method suggests potential applications across different scenarios and fraud types.

The findings underscore the importance of implementing advanced characterization and classification techniques in real smart-grid scenarios to detect fraud, potentially leading to significant cost reductions effectively. Applying these techniques could translate to real-world savings worth billions of dollars, assisting electricity suppliers in pinpointing the location, timing, and nature of frauds.

6.1 Future Works

The findings of this thesis lay a foundation for future academic inquiries and practical implementations in the field of electric power system fraud detection. The successful application of advanced characterization and classification techniques, as demonstrated by HybridForest, opens new avenues for research, particularly in enhancing fraud detection methodologies and developing cost-effective solutions for electricity suppliers.

Future works may explore the scalability of the proposed models across different datasets and real-world scenarios, the integration of additional data sources for even more nuanced fraud detection, and the application of these models in other sectors prone to fraudulent activities. Additionally, the continued refinement of TSC classifiers and ensemble learning methods could further improve the accuracy and efficiency of NTL detection, potentially leading to significant cost savings and operational efficiencies for electricity suppliers worldwide.

This thesis contributes to the academic discourse on fraud detection in electric power systems and provides practical insights that could significantly impact the management and operation of Smart Grids. By addressing the complexities of fraud detection with innovative methodologies, this work paves the way for more secure, efficient, and reliable electricity distribution networks.

6.2 Published Works

Part of the results of this proposal is already published in journals and conferences.

1. **BASTOS, LUCAS**; PFEIFF, G.; OLIVEIRA, R.; OLIVEIRA, H.; TOSTES, M. E.; ZEADALLY, S.; CERQUEIRA, E.; ROSARIO, D. . "Data-oriented Ensemble Predictor Based on Time Series Classifiers for Fraud Detection." In: *Electric Power Systems Research, 2022. Electric Power Systems Research, 2022.*
2. **BASTOS, LUCAS**; MARTINS, B.; MEDEIROS, I.; ROSARIO, D.; AQUINO, A. L.; CERQUEIRA, E. . "Energy Frauds Characterization based on Information Theory Quantifiers." In: *IWCMC, 2023. IWCMC, 2023.*
3. **BASTOS, LUCAS**; MARTINS, B. ; SANTOS, H. ; MEDEIROS, I. ; EUGENIO, P. ; MARQUES, L. ; ROSARIO, D. ; NOGUEIRA, E. ; CERQUEIRA, E. ; KREUTZ, M. ; NETO, A. . "Predictive Fraud Detection: An Intelligent Method for Internet of Smart Grid Things Systems". In: *JISA, 2023. JISA, 2023.*

Additional publications not necessarily related to the purpose of this research proposal were submitted and accepted at the following places:

1. **BASTOS, L.**; ROSARIO, D.; CERQUEIRA, E.; SANTOS, A.; NOGUEIRA, M. “Filtering Parameters Selection Method and Peaks Extraction for ECG and PPG Signals.“ In: *2019 IEEE Latin American Conference on Communications (LATINCOM), 2019, Salvador.* 2019 IEEE Latin-American Conference on Communications (LATINCOM), 2019. p. 1.
2. **BASTOS, LUCAS**; TAVARES, THAIS ; ROSARIO, DENIS ; CERQUEIRA, EDUARDO ; SANTOS, ALDRI ; NOGUEIRA, MICHELE . Double Authentication Model based on PPG and ECG Signals. In: *2020 International Wireless Communications and Mobile Computing (IWCMC)*, 2020, Limassol. 2020 International Wireless Communications and Mobile Computing (IWCMC), 2020. p. 601.
3. **BASTOS, LUCAS**; CREMONEZI, BRUNO ; TAVARES, THAIS ; ROSARIO, DENIS ; CERQUEIRA, EDUARDO ; SANTOS, ALDRI . “Smart Human Identification System Based on PPG and ECG Signals in Wearable Devices“. In: *2021 International Wireless Communications and Mobile Computing (IWCMC), 2021*, Harbin City. 2021 International Wireless Communications and Mobile Computing (IWCMC), 2021. p. 347.

4. **BASTOS, L. L.**; MARTINS, B. ; MEDEIROS, I. ; NETO, A. ; ZEADALLY, S. ; ROSARIO, D. ; CERQUEIRA, E. . "Ensemble Learning Method for Human Identification in Wearable Devices". In: *IWCMC, 2022, Dubrovnik. IWCMC 2022, 2022.*

5. **BASTOS, L.** ; SANTOS, C. ; MEDEIROS, I. ; FREITAS, I. ; ROSÁRIO, D. ; SERUFFO, M ; CERQUEIRA, E. . A Gamification and Biofeedback-based Serious Game for Adherence to Physical Activity. In: 2023 International Wireless Communications and Mobile Computing (IWCMC), 2023, Marrakesh. 2023 International Wireless Communications and Mobile Computing (IWCMC), 2023. p. 363.

6. Modesto, W.; **BASTOS, LUCAS**; NETO, A.; ROSARIO, D.; CERQUEIRA, E. . "Towards Automating the Integration of Legacy IEDs into Edge-Supported Internet of Smart Grid Things." In: *JISA, 2022. JOURNAL OF INTERNET SERVICES AND APPLICATIONS, 2022. v. 13.*

7. ARAUJO, F. ; **BASTOS, LUCAS** ; MEDEIROS, I. ; ROSSO, O. A. ; AQUINO, A. L. ; ROSARIO, D. ; CERQUEIRA, E. . "Characterization of human mobility based on Information Theory quantifiers". In: *Physica A: Statistical Mechanics and its Applications, 2023.* Physica A: Statistical Mechanics and its Applications, 2023.

8. SOARES, L. R. ; **BASTOS, LUCAS** ; MARTINS, B. ; MEDEIROS, I. ; ROSARIO, D. ; NOBRE, J. ; CERQUEIRA, EDUARDO . A Continuous Heart-Based Biometric Authentication for Healthcare Internet of Things. In: SBSeg, 2023. SBSeg, 2023.

9. MARQUES, L.; EUGENIO, P.; **BASTOS, LUCAS**; SANTOS, H.; ROSARIO, D.; NOGUEIRA, E.; CERQUEIRA, E.; KREUTZ, M.; NETO, A. . "Analysis of Electrical Signals by Machine Learning for Classification of Individualized Electronics in the Internet of Smart Grid Things (IoSGT) architecture." In: *JISA, 2023. JISA, 2023.*

References

[1] S. K. Gunturi and D. Sarkar, "Ensemble machine learning models for the detection of energy theft," *Electric Power Systems Research*, vol. 192, p. 106904, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378779620307021

[2] Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of smart meter data analytics: Applications, methodologies, and challenges," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 3125–3148, 2019.

[3] 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.

[4] L. A. P. Júnior, C. C. O. Ramos, D. Rodrigues, D. R. Pereira, A. N. de Souza, K. A. P. da Costa, and J. P. Papa, "Unsupervised non-technical losses identification through optimum-path forest," *Electric Power Systems Research*, vol. 140, pp. 413–423, 2016.

[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. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364032121004937

[6] 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.

[7] C. C. Ramos, D. Rodrigues, A. N. de Souza, and J. P. Papa, "On the study of commercial losses in brazil: a binary black hole algorithm for theft characterization," *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 676–683, 2018.

[8] 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, 2017.

[9] 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. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378779621003965

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

[11] P. Jokar, N. Arianpoo, and V. C. M. Leung, "Electricity theft detection in ami using customers' consumption patterns," *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 216–226, 2016.

[12] M. Alamaniotis and N. Gatsis, "Evolutionary multi-objective cost and privacy driven load morphing in smart electricity grid partition," *Energies*, vol. 12, no. 13, p. 2470, 2019.

[13] A. L. Aquino, H. S. Ramos, A. C. Frery, L. P. Viana, T. S. Cavalcante, and O. A. Rosso, "Characterization of electric load with information theory quantifiers," *Physica A: Statistical Mechanics and its Applications*, vol. 465, pp. 277–284, 2017.

[14] C. G. S. Freitas, A. L. L. Aquino, H. S. Ramos, A. C. Frery, and O. A. Rosso, "A detailed characterization of complex networks using information theory," *Scientific Reports*, vol. 9, no. 1, p. 16689, 2019.

[15] C. Bandt and B. Pompe, "Permutation entropy: a natural complexity measure for time series," *Physical review letters*, vol. 88, no. 17, p. 174102, 2002.

[16] M. Martin, A. Plastino, and O. Rosso, "Generalized statistical complexity measures: Geometrical and analytical properties," *Physica A: Statistical Mechanics and its Applications*, vol. 369, no. 2, pp. 439–462, 2006.

[17] I. S. S. D. Archive, "Commission for energy regulation (cer) smart metering project - electricity customer behaviour trial, 2009-2010 [dataset]," 2012. [Online]. Available: https://www.ucd.ie/issda/data/commissionforenergyregulationcer/

[18] S. K. Gunturi and D. Sarkar, "Ensemble machine learning models for the detection of energy theft," *Electric Power Systems Research*, vol. 192, p. 106904, 2021.

[19] F. Araujo, F. Araújo, K. Machado, D. Rosário, E. Cerqueira, and L. Villas, "Ensemble mobility predictor based on random forest and markovian property using lbsn data," *Journal of Internet Services and Applications*, vol. 11, 12 2020.

[20] C. H. Lubba, S. S. Sethi, P. Knaute, S. R. Schultz, B. D. Fulcher, and N. S. Jones, "catch22: Canonical time-series characteristics," *Data Mining and Knowledge Discovery*, vol. 33, no. 6, pp. 1821–1852, 2019.

[21] P. Schäfer and U. Leser, "Fast and accurate time series classification with weasel," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 637–646.

[22] H. Deng, G. Runger, E. Tuv, and M. Vladimir, "A time series forest for classification and feature extraction," *Information Sciences*, vol. 239, pp. 142–153, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025513001473

[23] V. Mahato, M. O'Reilly, and P. Cunningham, "A comparison of k-nn methods for time series classification and regression." in *AICS*, 2018, pp. 102–113.

[24] A. Dempster, F. Petitjean, and G. I. Webb, "Rocket: exceptionally fast and accurate time series classification using random convolutional kernels," *Data Mining and Knowledge Discovery*, vol. 34, no. 5, pp. 1454–1495, 2020.

[25] F. Araújo, D. Rosário, K. Machado, E. Cerqueira, and L. Villas, "Temmus: A mobility predictor based on temporal markov model with user similarity," in *Anais do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*. Porto Alegre, RS, Brasil: SBC, 2019, pp. 594–607. [Online]. Available: https://sol.sbc.org.br/index.php/sbrc/article/view/7389

[26] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition*. O'Reilly Media, Inc., 2019.

[27] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[28] L. A. Passos Júnior, C. C. Oba Ramos, D. Rodrigues, D. R. Pereira, A. N. de Souza, K. A. Pontara da Costa, and J. P. Papa, "Unsupervised non-technical losses identification through optimum-path forest," *Electric Power Systems Research*, vol. 140, pp. 413–423, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378779616302085

[29] S. Barja-Martinez, M. Aragüés-Peñalba, Íngrid Munné-Collado, P. Lloret-Gallego, E. Bullich-Massagué, and R. Villafafila-Robles, "Artificial intelligence techniques for enabling big data services in distribution networks: A review," *Renewable and Sustainable Energy Reviews*, vol. 150, p. 111459, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364032121007413

[30] 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.

[31] L. Bastos, G. Pfeiff, R. Oliveira, H. Oliveira, M. E. Tostes, S. Zeadally, E. Cerqueira, and D. Rosário, "Data-oriented ensemble predictor based on time series classifiers for fraud detection," *Electric Power Systems Research*, vol. 223, p. 109547, 2023.

[32] 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.

[33] G. M. Messinis, A. E. Rigas, and N. D. Hatziargyriou, "A hybrid method for non-technical loss detection in smart distribution grids," *IEEE Transactions on Smart Grid*, vol. 10, no. 6, pp. 6080–6091, 2019.

[34] M. A. de Souza, J. L. Pereira, G. d. O. Alves, B. C. de Oliveira, I. D. Melo, and P. A. Garcia, "Detection and identification of energy theft in advanced metering infrastructures," *Electric Power Systems Research*, vol. 182, p. 106258, 2020.

[35] S. Aoufi, A. Derhab, M. Guerroumi, H. Guemmouma, and H. Lazali, “Lite-fort: Lightweight three-stage energy theft detection based on time series forecasting of consumption patterns,” *Electric Power Systems Research*, vol. 225, p. 109840, 2023.

[36] R. K. Ahir and B. Chakraborty, “Pattern-based and context-aware electricity theft detection in smart grid,” *Sustainable Energy, Grids and Networks*, vol. 32, p. 100833, 2022.

[37] M. N. Hasan, R. N. Toma, A.-A. Nahid, M. M. Islam, and J.-M. Kim, “Electricity theft detection in smart grid systems: A cnn-lstm based approach,” *Energies*, vol. 12, no. 17, p. 3310, 2019.

[38] M. Adeli, M. Hajatipour, M. J. Yazdanpanah, H. Hashemi-Dezaki, and M. Shafieirad, “Optimized cyber-attack detection method of power systems using sliding mode observer,” *Electric Power Systems Research*, vol. 205, p. 107745, 2022.

[39] S. Zidi, A. Mihoub, S. M. Qaisar, M. Krichen, and Q. A. Al-Haija, “Theft detection dataset for benchmarking and machine learning based classification in a smart grid environment,” *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 1, pp. 13–25, 2023.

[40] R. Xia, Y. Gao, Y. Zhu, D. Gu, and J. Wang, “An attention-based wide and deep cnn with dilated convolutions for detecting electricity theft considering imbalanced data,” *Electric Power Systems Research*, vol. 214, p. 108886, 2023.

[41] 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.

[42] 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. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0142061516316386

[43] M. Zanetti, E. Jamhour, M. Pellenz, M. Penna, V. Zambenedetti, and I. Chueiri, “A tunable fraud detection system for advanced metering infrastructure using short-lived patterns,” *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 830–840, 2019.

[44] G. M. Messinis, A. E. Rigas, and N. D. Hatziargyriou, “A hybrid method for non-technical loss detection in smart distribution grids,” *IEEE Transactions on Smart Grid*, vol. 10, no. 6, pp. 6080–6091, 2019.

[45] S.-C. Yip, W.-N. Tan, C. Tan, M.-T. Gan, and K. Wong, “An anomaly detection framework for identifying energy theft and defective meters in smart grids,” *International Journal of Electrical Power Energy Systems*, vol. 101, pp. 189–203, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0142061517318719

[46] 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), 2016, pp. 407–418.

[47] P. W. Lamberti, M. Martin, A. Plastino, and O. A. Rosso, "Intensive entropic non-triviality measure," *Physica A: Statistical Mechanics and its Applications*, vol. 334, no. 1-2, pp. 119–131, 2004.

[48] R. López-Ruiz, H. L. Mancini, and X. Calbet, "A statistical measure of complexity," *Physics Letters A*, vol. 209, no. 5-6, pp. 321–326, 1995.

[49] F. Olivares, L. Souza, W. Legnani, and O. A. Rosso, "Informational time causal planes: A tool for chaotic map dynamic visualization," in *Nonlinear Systems-Theoretical Aspects and Recent Applications*. IntechOpen, 2019.