American Journal of Engineering and Applied Sciences

Research Article

Optimization of Non-Technical Electricity Loss Detection Using Artificial Neural Networks: Case of Cameroon Distribution Network

Lekini Nkodo Claude Bernard[1], Bell Serge Samuel[1], Nyemb Nsoga Valjacques[1] and Nzotcha Urbain[2,3]

[1]*Department of Renewable Energy, Institute of Wood Technology, University of Yaounde I, Cameroon*
[2]*Department of Mechanical, National Advanced School of Engineering of Yaoundé, University of Yaoundé I, Cameroun*
[3]*Forschungszentrum Jülich GmbH, Institute of Energy and Climate Research Fundamental Electrochemistry (IEK-9), Wilhelm-Johnen Straße, 52428 Jülich, Germany*

Article history
Received: 12-01-2026
Revised: 19-05-2026
Accepted: 04-06-2026

Corresponding Author:
Lekini Nkodo Claude Bernard
Electrical, Department of Renewable Energy of the Institute of Wood Technology, University of Yaounde I, Cameroon
Email: claudelekini@gmail.com

Abstract: Non-Technical Losses (NTL) represent a critical challenge for power utilities, particularly in developing countries. This paper proposes an unsupervised deep learning approach based on an autoencoder to detect abnormal electricity consumption patterns in the Cameroonian distribution network. The model learns normal customer behavior using historical consumption data and identifies anomalies through reconstruction error analysis. Statistical thresholds are applied to classify customers as normal, suspicious, or fraudulent. Experimental results show that the proposed method achieves high detection reliability while reducing inspection costs. The approach provides a scalable and practical solution for utility companies seeking to enhance revenue protection and energy security.

Keywords: Non-Technical Losses, Autoencoder, Fraud Detection, Customer Behavior, Data, Load Profile, Irregularities, Prediction

Introduction

In electrical distribution systems, Non-Technical Losses (NTL) represent a major challenge for utilities. Unlike technical losses, which arise from physical properties of the network, NTL are primarily caused by fraudulent activities, administrative failures, or metering errors. These losses are particularly critical because they are invisible, unpredictable, and difficult to quantify, making their detection a highly complex and costlay process.

In Cameroon, this issue has reached alarming proportions. According to Lekini et al. (2017), NTL accounts for approximately 30% of the energy injected into the grid, resulting in annual financial losses exceeding 60 billion FCFA. Such losses undermine the economic sustainability of utilities, reduce the ability to invest in modern infrastructures, and negatively impact service quality and energy security.

The main causes of NTL include:

- Fraudulent activities: Illegal connections, meter tampering, and technical sabotage
- Administrative failures: Incorrect billing, post-termination consumption, and inaccurate meter index recording
- Lack of systematic monitoring: Absence of regular inspections of customer installations and metering equipment
- Rapid and uncontrolled urbanization: Development of informal settlements complicating network supervision

These irregularities not only create significant financial losses but also erode customer trust, hinder social equity, and compromise national energy transition goals.

Traditional detection methods, such as physical audits and manual meter inspections, have proven inefficient, costly, and limited in scope. This situation calls for innovative solutions capable of processing large-scale data and identifying abnormal consumption patterns with high accuracy. Artificial Intelligence (AI), particularly Artificial Neural Networks (ANNs), provides a promising alternative by enabling predictive and proactive detection of NTL. Artificial Neural Networks originate from the pioneering work of McCulloch and Pitts (1943), later formalized through the perceptron model proposed by Rosenblatt (1958), and significantly improved with the backpropagation algorithm introduced by Rumelhart et al. (1986).

Therefore, the key research question can be formulated as:

How can an Artificial Neural Network-based approach be designed and implemented to effectively detect Non-Technical Losses in Cameroon's power distribution networks, while considering local technical, economic, and social constraints?

The general objective is to develop and evaluate an intelligent ANN-based model for the detection and reduction of Non-Technical Losses in Cameroon's power distribution system, with the ultimate goal of improving operational efficiency, financial performance, and energy security.

Specific Objectives ARE

1. Analyze the root causes and characteristics of Non-Technical Losses within the Cameroonian power sector using historical consumption and billing data
2. Design a predictive ANN-based model capable of learning normal consumption patterns and detecting anomalies indicative of fraud or operational errors
3. Evaluate model performance using key metrics such as accuracy, precision, recall, and F1-score, and compare the results with conventional detection approaches
4. Propose an operational integration framework for the ANN model within existing utility systems, considering infrastructure, cost, and social acceptance constraints
5. Provide strategic and technical recommendations to optimize NTL detection in Cameroon's electricity distribution network

This research addresses a critical challenge for energy sustainability and financial viability in Cameroon's electricity sector. With NTL accounting for nearly 30% of distributed energy, utilities face substantial revenue losses, which limit their capacity to invest in network modernization, smart metering, and grid reliability improvements.

Current detection strategies, such as manual inspections and periodic audits, suffer from multiple drawbacks:

- High operational costs and limited coverage
- Delayed detection of fraudulent activities, leading to prolonged revenue leakage
- Inability to process large-scale, high-dimensional data generated by millions of customers

Conversely, Artificial Neural Networks offer significant advantages:

- Ability to process massive and heterogeneous datasets (billing history, meter readings, geolocation data) in near real-time
- Capability to identify complex, nonlinear anomalies that traditional statistical methods fail to capture
- Cost reduction and efficiency improvement in fraud detection and revenue assurance

Furthermore, this research aligns with national priorities for modernizing the power sector, promoting digital transformation, and supporting energy transition goals in sub-Saharan Africa.

Related Work and Literature Review

Over the past decade, the detection and mitigation of Non-Technical Losses (NTL), particularly electricity theft, has gained significant attention due to its economic and operational impact on power distribution systems. This section reviews major contributions from 2016 to 2025, highlighting methodologies, trends, and research gaps.

Carr and Thomson (2022) presented a comprehensive scope review of non-technical losses, clarifying the definitions, causes, and socio-economic implications of electricity theft. Their work stressed that technical solutions, such as machine learning, must be complemented by regulatory frameworks and consumer awareness programs, particularly in developing countries where theft is prevalent.

Saeed et al. (2020) conducted a systematic review of NTL detection techniques, cataloging machine learning and deep learning algorithms (e.g., decision trees, SVMs, neural networks) and discussing persistent challenges like data imbalance and cross-country model transferability. They also provided recommendations for deployment-ready detection systems, making their study a cornerstone for utility applications. Earlier studies by Depuru et al. (2011); Monedero et al. (2006); Leite and Mantovani (2018); Buzau et al. (2019) demonstrated the effectiveness of machine learning techniques for electricity theft detection using smart meter data and distribution network analytics.

Earlier, Viegas et al. (2017) focused on data analytics solutions for fraud detection in power systems, proposing effective feature engineering methods and hybrid models that combine rule-based systems with machine learning. Their

findings suggested that such hybridization can significantly reduce false positives, a common issue in large-scale NTL detection.

In recent years, the integration of smart meters has transformed theft detection methodologies. Xia et al. (2022) reviewed smart meter-based techniques, emphasizing deep learning architectures, multi-sensor data fusion, and cybersecurity against adversarial attacks. Similarly, Althobaiti et al. (2021) offered a data-driven survey linking detection techniques to smart grid functional layers and fraud strategies. They discussed supervised, semi-supervised, and unsupervised approaches while introducing concerns regarding data poisoning and privacy.

The evolution toward deep learning solutions is further reflected in Guarda et al. (2023), who explored non-hardware-based methods, primarily using historical consumption data. They evaluated rule-based, statistical, and AI-driven strategies, emphasizing data preprocessing, class balancing, and advanced performance metrics (AUC, F1-score). Another study by Badr et al. (2023) systematically reviewed feature extraction methods and neural network models, highlighting challenges such as model generalization across utilities and lack of open datasets.

Beyond technical factors, Stracqualursi et al. (2023) examined the socio-economic dimension of electricity theft and how these factors influence AI-based detection systems. They proposed cost-benefit frameworks to guide decision-making for distribution system operators (DSOs), particularly relevant for resource-constrained environments like Africa.

Recent advances (2019–2024) have also focused on addressing data imbalance and enhancing generalization through data augmentation and generative models (GANs, VAEs). A notable Gong et al. (2020) documented the application of these models for representation learning and robust fraud detection. Similarly, Kgaphola et al. (2024) provided a practical perspective, detailing AMI analytics, embedded AI systems, and field inspection optimization while identifying gaps in interoperability and ethical considerations.

Materials and Methods

Materials

The database provides a monthly overview of individual electricity consumption, combining technical, administrative, and behavioral variables, thus offering a robust analytical framework for studying irregularities and consumption patterns within the distribution network. The quality and representativeness of the data are fundamental to the effectiveness of an autoencoder. Given the unsupervised nature of the model, the data used for training must accurately reflect the normal behaviour of the system under study. Any excessive noise, outliers, or unrepresentative data can affect the learning of the latent structure and degrade reconstruction performance. The data must be sufficiently large and diverse to capture the natural variability of the phenomenon being analysed. Rigorous pre-processing is essential, including cleaning missing values, removing or processing outliers, and normalising or standardising variables. This step ensures stable convergence of the model and prevents a dominant variable from unduly influencing learning.

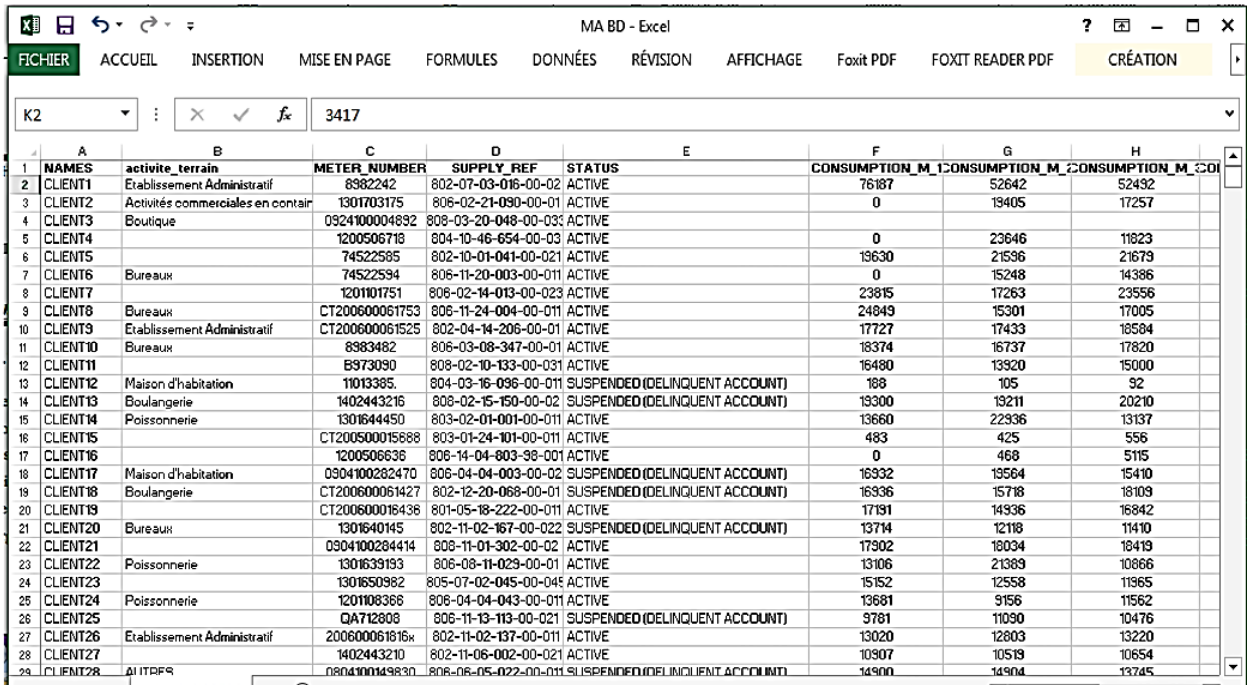
Concerning our model, we have a database of 61,782 customers as presented in Fig. 1. Data comes from electricity consumption records provided by Cameroon's national electricity distribution company, ENEO Cameroon S.A. It was extracted from the company's commercial and technical information system, which records periodic meter readings from subscribers. The database covers a period of 24 consecutive months, allowing for a longitudinal analysis of consumption patterns. The data concerns active customers of the distribution network, belonging to different socio-economic segments (administrative, commercial, residential, etc.), which guarantees a diversity of energy profiles studied. The extraction was carried out in accordance with the operator's internal data management procedures, and the information was anonymized to preserve the confidentiality of subscribers.

Methods

Neural networks are machine learning models that mimic the complex functioning of the human brain. These models consist of interconnected nodes or neurons that process data, learn patterns, and enable tasks such as pattern recognition and decision-making (Goodfellow et al., 2016). Neural networks are capable of learning and identifying patterns directly from data without pre-defined rules. Classical neural network theory has been extensively described by Haykin (1999), while Goodfellow et al. (2016); Chollet (2017) provide modern deep learning implementations.

Our methodology will be based on the autoencoder, which is an unsupervised neural network that learns to reconstruct its input data after compressing it into a lower-dimensional latent space. It consists of an encoder, which reduces the dimension of the data, and a decoder, which reconstructs it. Learning consists of minimising the

reconstruction error between the input and the output. Autoencoders are used for dimension reduction, denoising, data compression, and anomaly detection (fraud, defects, non-technical losses). Deep learning architectures have evolved through recurrent neural networks (Elman, 1990), Long Short-Term Memory networks (Hochreiter and Schmidhuber, 1997), convolutional neural networks (LeCun et al., 1998), and more recently Transformer architectures (Vaswani et al., 2017). Although these architectures are powerful, autoencoders remain particularly suitable for unsupervised anomaly detection.



	A	B	C	D	E	F	G	H
1	NAMES	activite_terrain	METER_NUMBER	SUPPLY_REF	STATUS	CONSUMPTION_M_1	CONSUMPTION_M_2	CONSUMPTION_M_3
2	CLIENT1	Etablissement Administratif	8982242	802-07-03-016-00-02	ACTIVE	76187	52642	52492
3	CLIENT2	Activités commerciales en contair	1301703175	806-02-21-090-00-01	ACTIVE	0	19405	17257
4	CLIENT3	Boutique	0924100004892	808-03-20-048-00-03	ACTIVE			
5	CLIENT4		1200506718	804-10-46-654-00-03	ACTIVE	0	23646	11823
6	CLIENT5		74522585	802-10-01-041-00-021	ACTIVE	13630	21596	21679
7	CLIENT6	Bureaux	74522594	806-11-20-003-00-011	ACTIVE	0	15248	14386
8	CLIENT7		1201101751	806-02-14-013-00-023	ACTIVE	23815	17263	23556
9	CLIENT8	Bureaux	CT200600061753	806-11-24-004-00-011	ACTIVE	24849	15301	17005
10	CLIENT9	Etablissement Administratif	CT200600061525	802-04-14-206-00-01	ACTIVE	17727	17433	18584
11	CLIENT10	Bureaux	8983482	806-03-08-347-00-01	ACTIVE	18374	16737	17820
12	CLIENT11		B973090	808-02-10-133-00-031	ACTIVE	16480	13920	15000
13	CLIENT12	Maison d'habitation	11013385.	804-03-16-096-00-011	SUSPENDED (DELINQUENT ACCOUNT)	188	105	92
14	CLIENT13	Boulangerie	1402443216	808-02-15-150-00-02	SUSPENDED (DELINQUENT ACCOUNT)	19300	19211	20210
15	CLIENT14	Poissonnerie	1301644450	803-02-01-001-00-011	ACTIVE	13660	22936	13137
16	CLIENT15		CT200500015688	803-01-24-101-00-011	ACTIVE	483	425	556
17	CLIENT16		1200506636	806-14-04-803-98-001	ACTIVE	0	468	5115
18	CLIENT17	Maison d'habitation	0904100282470	806-04-04-003-00-02	SUSPENDED (DELINQUENT ACCOUNT)	16932	13564	15410
19	CLIENT18	Boulangerie	CT200600061427	802-12-20-068-00-01	SUSPENDED (DELINQUENT ACCOUNT)	16936	15718	18109
20	CLIENT19		CT200600016436	801-05-18-222-00-011	ACTIVE	17191	14936	16842
21	CLIENT20	Bureaux	1301640145	802-11-02-167-00-022	SUSPENDED (DELINQUENT ACCOUNT)	13714	12118	11410
22	CLIENT21		0904100284414	808-11-01-302-00-02	ACTIVE	17902	18034	18419
23	CLIENT22	Poissonnerie	1301639193	806-08-11-029-00-01	ACTIVE	13106	21389	10866
24	CLIENT23		1301650982	805-07-02-045-00-045	ACTIVE	15152	12558	11965
25	CLIENT24	Poissonnerie	1201108368	806-04-04-043-00-011	ACTIVE	13681	9156	11562
26	CLIENT25		QA712808	806-11-13-113-00-021	SUSPENDED (DELINQUENT ACCOUNT)	9781	11090	10476
27	CLIENT26	Etablissement Administratif	200600061816x	802-11-02-137-00-011	ACTIVE	13020	12803	13220
28	CLIENT27		1402443210	802-11-06-002-00-021	ACTIVE	10907	10519	10654
29	CLIENT28	AUTRES	0804100149830	806-06-05-022-00-011	SUSPENDED (DELINQUENT ACCOUNT)	14900	14904	13745

Fig. 1: Customer consumption data from ENEO

In our case, fraud detection is treated as an unsupervised anomaly detection problem. The methodology follow a 'compression-reconstruction' process:

- ✓ Learning normal behaviour: The model is trained to compress the input data into a reduced latent space (bottleneck), then to reconstruct it as faithfully as possible
- ✓ Bottleneck: By forcing the data through a small number of neurons (8 latent neurons), the model learns only the essential characteristics of normal consumption
- ✓ Reconstruction error (MSE): Detection reliability is based on Mean Square Error (MSE). A 'normal' customer will have a low error because the model knows their profile. A "fraudulent" or 'suspicious' customer will have a high MSE because their atypical behaviour cannot be accurately reconstructed by the model
- ✓ Adaptive thresholds: Classification is not based on pre-existing labels but on statistical percentiles of the error distribution (93rd percentile for suspects and 97th for fraudsters)

Implementation of Neural Network

Implementation starts with problem definition and data collection, followed by data preprocessing, including cleaning and normalization. An autoencoder architecture composed of an encoder, a latent space, and a decoder is then designed. Appropriate activation functions and a reconstruction loss function are selected before training the model in an unsupervised manner. Model performance is evaluated using the reconstruction error. Finally, the learned latent representation is exploited for downstream tasks such as dimensionality reduction, clustering, or anomaly detection, and the model is optimized and deployed as follows: The steps are presented in Fig. 2.

DOI: 10.3844/ajeassp.2026.117.132

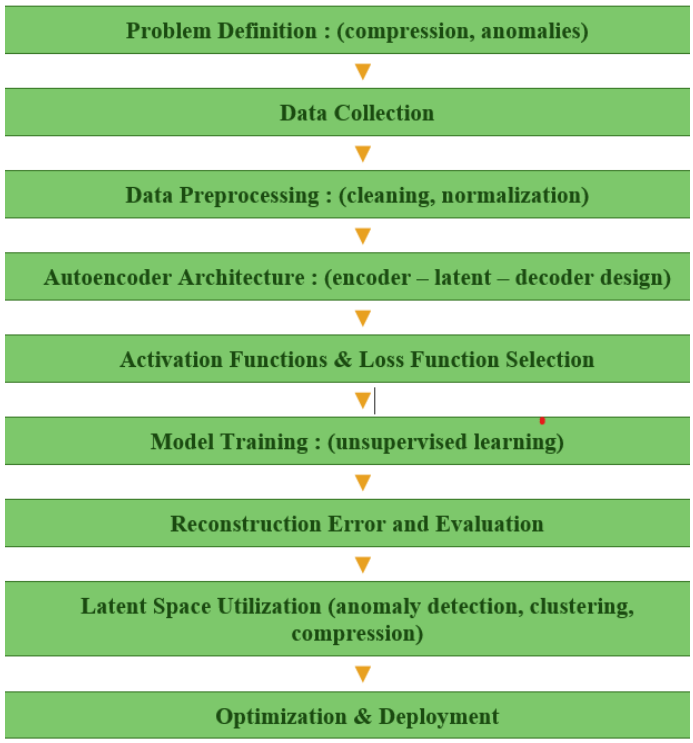


Fig. 2: Implémentation diagram

Data Loading

The dataset MA BD.xlsx, comprising 24 months of customer consumption data, was imported and preprocessed for analysis. Key variables included customer identifiers, consumption, activity type, and contractual status, which are critical for detecting anomalies such as sudden drops in consumption or atypical behavioral patterns. Data manipulation was performed using pandas and numpy, with openpyxl employed for specific Excel formats. Consumption values were cast as numeric types, and date fields were standardized as datetime objects to ensure computational accuracy and temporal consistency.

Data Cleaning and Preprocessing

Data cleaning and preprocessing were conducted using a rigorous approach to ensure the reliability of the analysis. Redundant observations were removed, as well as customers with no consumption over the entire study period and those whose presence was not consistent throughout the 24 months, in order to preserve temporal homogeneity. Missing values were handled according to their nature: Median imputation was applied to numerical variables, interpolation was used for isolated gaps in consumption, and zero values were assigned in cases of confirmed non-consumption. Finally, an overall quality check was performed to correct inconsistencies and eliminate any remaining duplicates, thereby ensuring the integrity of the dataset.

Feature Engineering

Feature engineering was performed to enhance the dataset for modeling customer behavior. Derived variables included six-month rolling means to identify high-consumption customers, standard deviation and coefficient of variation to quantify stability and relative variability, and maximum/minimum values to capture consumption peaks. Temporal features such as consumption trends, number of zero-consumption months, and timing of abrupt increases or decreases were incorporated. Peak detection was applied to identify anomalies indicative of potential fraud, for example, sharp sequential drops in consumption that may suggest meter bypass or tampering.

Encoding and Normalization

Categorical variables were encoded to enable integration into the deep learning model, e.g., activite_terrain → 0, 1, 2… and status → ACTIVE = 0, SUSPENDED =1. Numerical features were normalized using Robust Scaler, which centers values around the median and scales by the interquartile range, providing robustness against outliers.

DOI: 10.3844/ajeassp.2026.117.132

Error Calculation

After training, all customers were passed through the autoencoder to obtain reconstructed consumption values. The reconstruction error was computed as the difference between actual and reconstructed consumption according to Equation (1):

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (X_i - X'_i) \quad (1)$$

Low errors indicate normal behavior, whereas high errors suggest anomalous or potentially fraudulent activity.

Threshold Determination

To categorize customers based on reconstruction error, the distribution of Mean Squared Errors (MSE) was analyzed. Thresholds were defined as follows: Errors ≤ 85th percentile indicate normal behavior, errors between the 85th and 95th percentiles indicate suspicious activity, and errors ≥ 95th percentile indicate likely fraudulent behavior. This percentile-based approach is robust, as it adapts automatically to the characteristics of the dataset.

Analysis and Visualization

Interactive dashboards were developed using charts such as pie, bar, line, and scatter plots to analyze customer behavior. Visualizations included the distribution of customers by category (normal, suspicious, fraudulent), six-month consumption trends, MSE error histograms, comparisons across activity types (e.g., barbershop, household, bar), and heatmaps of anomalous behavior. Data filtering was enabled by activity type, contractual status, geographic zone, and consumption level, facilitating targeted interventions by field and anti-fraud teams.

Hyperparameter Table

Hyperparameters play a decisive role in the performance and generalisation capacity of autoencoders. In our case, we will use the following hyperparameters in Table 1.

Table 1: Hyperparameter

Parameter	Specification
Number of layers	7 (Input, 3 Encoder, 2 Decoder, Output)
Nomber neurons	32 → 16 → 8 (Latent) → 16 → 32
Activation function	ReLU (hidden layers) et Linear (output layer)
Optimizer	Adam with a learning rate (LR) of 0.0005
Epochs	100 (with Early Stopping at a patience of 10)
Loss Function	MSE (Mean Squared Error)
Regularisation	L2 (0.01) et Dropout (0.4)
Batch size	512

Development Environment

According to the execution traces and system warnings present in the sources we have:

- ✓ Language: Python 3.10 (identified via the system paths \Python310\)
- ✓ Main libraries: TensorFlow and Keras
- ✓ Pre-processing tools: Scikit-learn (for RobustScaler, LabelEncoder, train_test_split)

```
pandas==2.1.1          # DataFrame management
numpy==1.24.3          # Numerical and vector calculations
scikit-learn==1.3.1    # Scalers and encoders
tensorflow==2.15.0     # For the Autoencoder model, the Keras models come from
```

TensorFlow

- ✓ Visualisation: Matplotlib and Seaborn, regardless of version

DOI: 10.3844/ajeassp.2026.117.132

Key Features of Our Model

Objectives

The main objectives are to develop a model based on unsupervised deep learning (autoencoder) to automatically detect anomalies in customer electricity consumption:

- Clean and transform monthly consumption data
- Extract characteristics describing each customer's energy behavior
- Build an Autoencoder capable of learning “normal” behavior from fraudulent or suspicious behavior
- Detect customers whose profiles show abnormal deviations
- Classify customers as: Normal – Suspicious – Fraudulent
- Generate analytical reports by activity or by area

Key Features

Advanced preprocessing of 24 months of consumption data:

- ✓ Calculation of descriptive statistics for each customer: mean, standard deviation, maximum, minimum, median, and coefficient of variation
- ✓ Detection of months with zero consumption or missing data, and identification of extreme variations and temporal trends
- ✓ Robust data cleaning and normalization to prepare the model

Deep Learning Autoencoder

- ✓ Symmetric neural network designed to reconstruct normal consumption profiles
- ✓ Data compression in a latent layer to extract meaningful representations
- ✓ Application of regularization and dropout to avoid overfitting
- ✓ Calculation of the Mean Square Error (MSE) between actual and reconstructed consumption
- ✓ Profiles with errors exceeding a certain statistical threshold are considered suspicious or fraudulent

Automatic Classification

- ✓ Assignment of a status to each customer according to the extent of the anomaly
- ✓ Normal: Consumption consistent with historical profile
- ✓ Suspicious: Moderate anomalies, to be monitored
- ✓ Fraudulent: Significant anomalies, intervention required

Detailed Analysis by Activity

- ✓ Segmentation of customers according to their type of activity (industries, shops, households, administrations, etc.)
- ✓ Assessment of the fraud rate by category to identify sectors at risk
- ✓ Fraud score out of 100
- ✓ Assign a fraud score to each customer based on the reconstruction error and statistical indicators
- ✓ Enables direct ranking of customers from safest to riskiest

Top 20 Potentially Fraudulent Customers

- ✓ List of highest-risk customers with details: Score, type of activity, status
- ✓ Facilitates rapid decision-making for control teams

Interactive Visualizations

- ✓ Error distribution: Visual identification of anomalies and thresholds
- ✓ Autoencoder learning curves: Monitoring of model convergence and performance
- ✓ Consumption vs. fraud score: Correlation between actual consumption and estimated risk
- ✓ Distribution by activity: Comparison of the number of fraud cases and the average score according to activity type

DOI: 10.3844/ajeassp.2026.117.132

Results and Discussion

Results

Graphics Interface

Web Dashboard For Visualization

The graphical interface implemented as part of our research is as follows in Fig. 3. This result indicates that the customer exhibits atypical consumption behaviour, which the autoencoder detects as difficult to reconstruct accurately. Although the anomalies observed are not significant enough to conclude that fraud has definitely occurred, they justify increased monitoring or even a targeted inspection to confirm or rule out a possible non-technical loss. The overall distribution shows that: The majority of customers are classified as normal (≈ 95%), a limited proportion are suspicious (≈ 4%), and a marginal fraction are fraudulent (≈ 1%). The customer under review, therefore, belongs to an intermediate zone, requiring special attention without being immediately considered fraudulent.

The monthly consumption profile Analysis shows that: Significant variations between months, high peaks (particularly around May and October), and sharp drops (February, March and August).

Statistical indicators confirm this variability: Average consumption: 185.3 kWh, Standard deviation: 51.1 kWh, indicating significant dispersion, Min: 100.7 kWh / Max: 280.7 kWh, reflecting a wide range of consumption. This high dispersion is atypical compared to normal profiles, which are generally more regular.

Automatic Classification

Here we can search by name, i.e., the user enters a customer's name, and the system displays their detailed information, a graph showing their monthly consumption, and the AI's decision: Normal, Suspicious, or Fraudulent. The user can also perform a random search by status, i.e., they select a status (Normal, Suspicious, or Fraudulent), and the system returns a customer matching that profile, along with their information and consumption graph, as presented in Fig. 4.

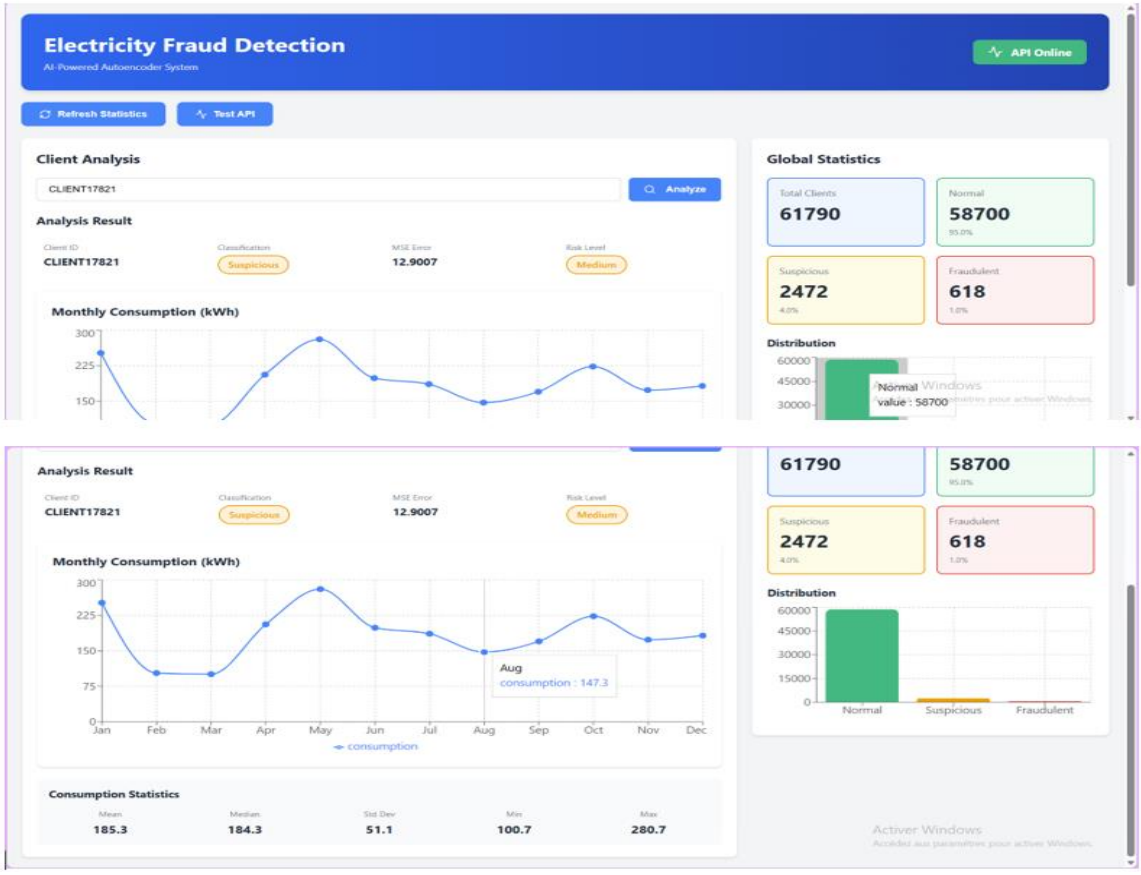


Fig. 3: Web dashboard for visualization

DOI: 10.3844/ajeassp.2026.117.132

Fig. 4: Detailed information of the Customer's

Top 20 Potentially Fraudulent Customers

As part of our research, an in-depth analysis of the data by our application identified a set of anomalies characteristic of potentially fraudulent behavior. The ranking below in Fig. 5 highlights the twenty customers most likely to be involved in irregular activities. This identification is an essential step in refining detection strategies and strengthening monitoring mechanisms.

Detailed Analysis By Activity

Here, the user selects a type of activity, and the system displays the detailed analysis corresponding to that activity, as shown in Figs. 6 and 7.

DOI: 10.3844/ajeassp.2026.117.132

	NAMES	METER_NUMBER	SUPPLY_REF	fraud_score
61644	CLIENT61645	XXXXXXXX_(A)	807-10-01-163-00-011	6151.042616
61475	CLIENT61476	EPESTIMMIMBO1	805-06-35-014-00-041	4479.477399
60735	CLIENT60736	1EPESTIME	801-07-01-008-00-021	3546.852892
60972	CLIENT60973	6715606	806-01-07-039-00-011	3102.997319
59095	CLIENT59096	EP20082706EST	804-01-01-859-01-011	2944.376327
58698	CLIENT58699	X04036586	803-07-13-015-00-011	2683.671823
60262	CLIENT60263	000000001	802-01-01-917-00-011	1661.004933
59784	CLIENT59785	2EPESTIME	808-02-01-002-00-031	1631.879372
60879	CLIENT60880	0904100284781	805-09-01-558-98-0101	707.876946
0	CLIENT1	8982242	802-07-03-016-00-02	592.609237
58693	CLIENT58694	CT200600061794	801-06-27-010-00-011	180.261183
228	CLIENT229	CT200500015764	802-05-01-001-00-031	160.667093
60039	CLIENT60040	75884070	807-02-09-121-00-011	145.965448
61196	CLIENT61197	0924100000489	807-20-01-816-98-001	79.036972
10899	CLIENT10900	04036562	801-03-29-009-00-011	32.878177
1424	CLIENT1425	0904100286841	802-13-05-005-00-071	32.289168
61050	CLIENT61051	0802100139589	803-16-15-787-00-011	32.138070
514	CLIENT515	1301701983	806-06-09-213-00-011	31.318971
59212	CLIENT59213	0000874	805-11-16-015-00-01	26.391737
1007	CLIENT1008	0904100180193	802-06-03	24.580237

Fig. 5: Top 20 Potentially fraudulent customers

Here, the user selects a type of activity, and the system displays the detailed analysis corresponding to that activity, as shown in Figs. 6 and 7. Fig. 7: Distribution of missing values before and after by customer.

Exploratory Data Analysis

Exploratory Data Analysis, which is an approach that uses visual and statistical methods to explore a dataset in order to summarize its main characteristics, discover patterns, identify anomalies, test hypotheses, and better understand the relationships between variables, prior to any formal modeling, while evaluating the quality of the data to determine the appropriate techniques, gave us the following result in Fig. 8.

Abnormal values were detected using IQR and Z-Score, and customers with zero consumption or ≥6 months at zero were identified. Abnormal values must be corrected or excluded if they result from technical errors (data entry, sensor), using methods such as Z-score or IQR. However, if they represent useful abnormal behavior (fraud, failure), they must be retained and processed using robust transformations or specialized models such as an Autoencoder, among others, as follows in Fig. 9.

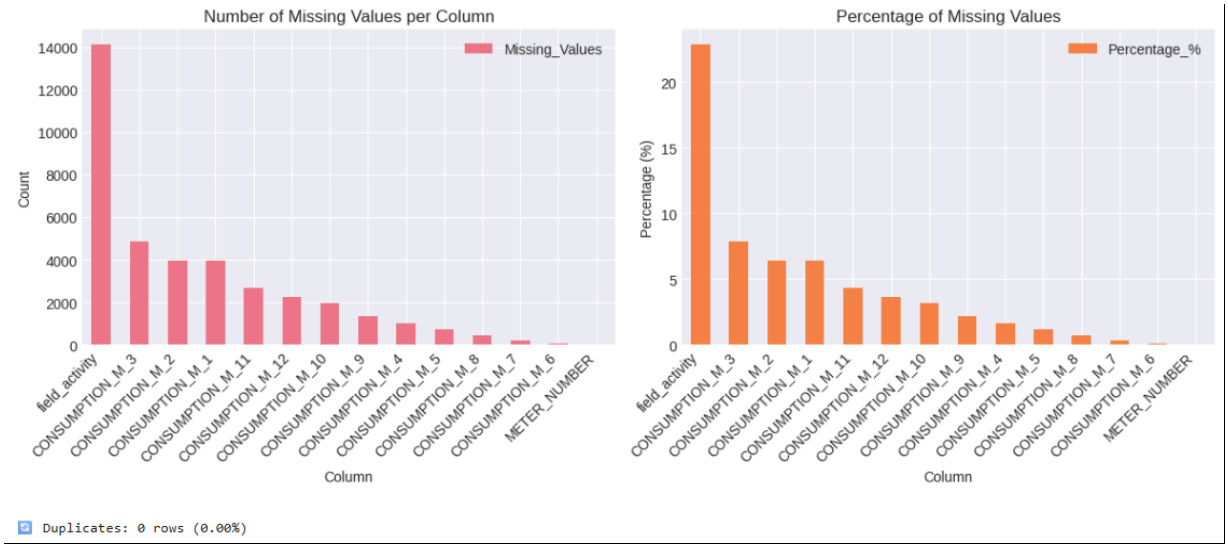


Fig. 6: Processing of missing values before and after

DOI: 10.3844/ajeassp.2026.117.132

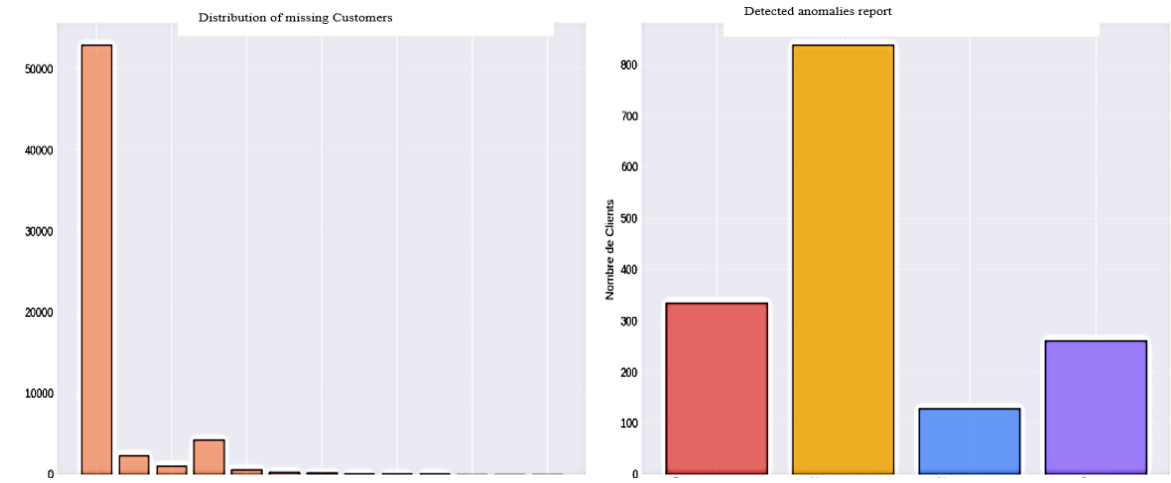


Fig. 7: Distribution of missing values before and after by customer

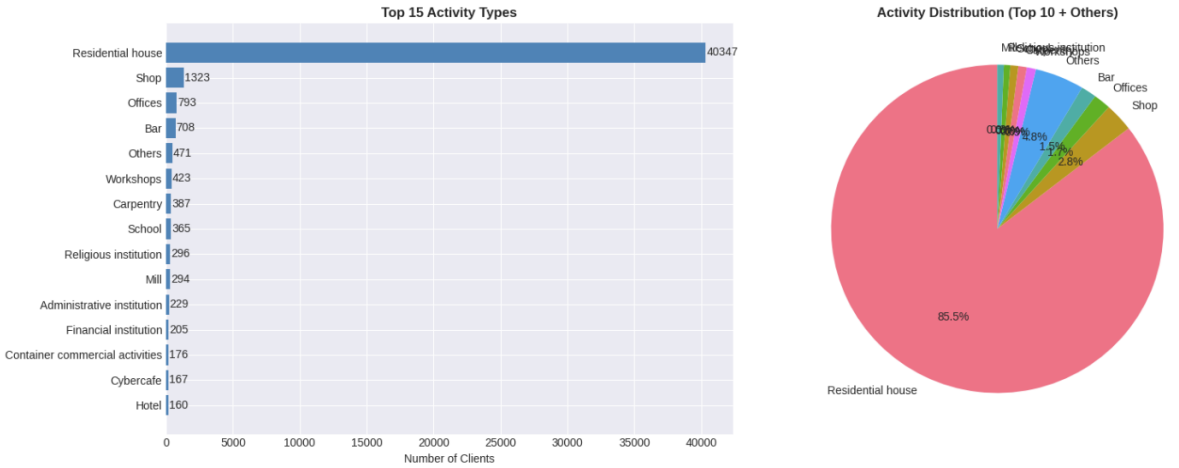


Fig. 8: Exploratory Data Analysis

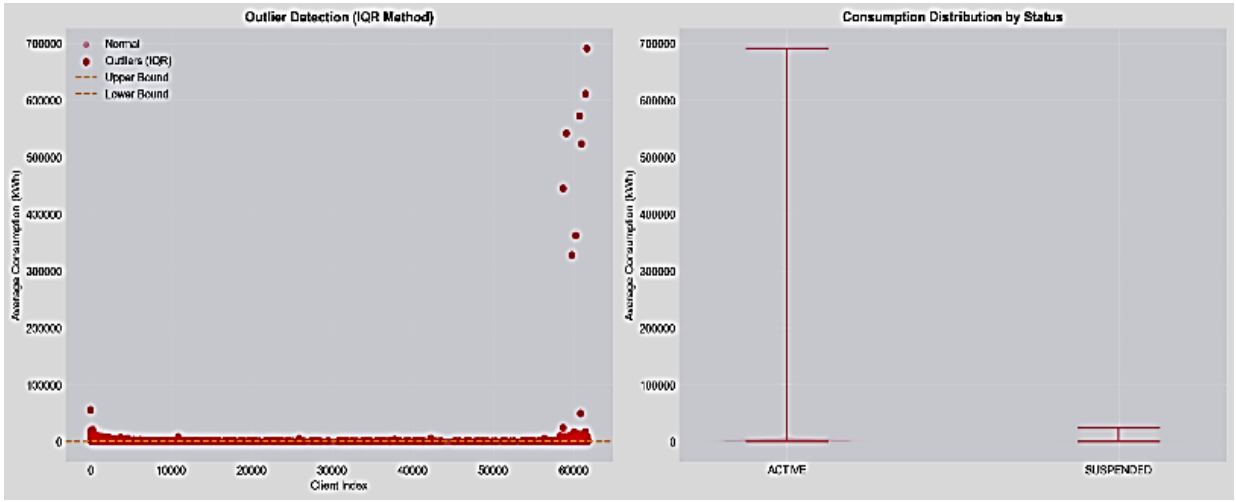


Fig. 9: Outlier distribution graph

DOI: 10.3844/ajeassp.2026.117.132

The activity analysis of average consumption and standard deviations for the top 10 activities was analyzed and visualized as shown in the following Fig. 10.

Anomalies include extreme variations, high CV, suspended accounts with high consumption, and sudden drops in consumption, as shown in the following Fig. 11.

The correlation matrix between the numerical variables presented in Fig. 12 was explored to identify significant correlations, and this matrix confirms that a feature selection step is necessary to eliminate redundancy in the volume measurements. For the Autoencoder, the objective is to provide it with variables that capture unique aspects of customer behavior.

Model Reliability

The fraud detection model is considered reliable because its methodology is rigorously aligned with the principles of unsupervised anomaly detection, based on the Measurement of Reconstruction Error (MSE).

Here are the detailed justifications for reliability.

Ability to Reconstruct Normal Behavior

Reliability is based on the intrinsic function of the Autoencoder, which is to Minimize Error (MSE) for training data considered "normal." Therefore, when the model attempts to reconstruct an atypical (potentially fraudulent) consumption profile, it fails and produces a significantly higher reconstruction error. This error is the key metric for the anomaly. For example, the at-risk customer CUSTOMER1 has a reconstruction error of 189617.114606 and a fraud score of 100.0, demonstrating the model's ability to isolate extreme behaviors.

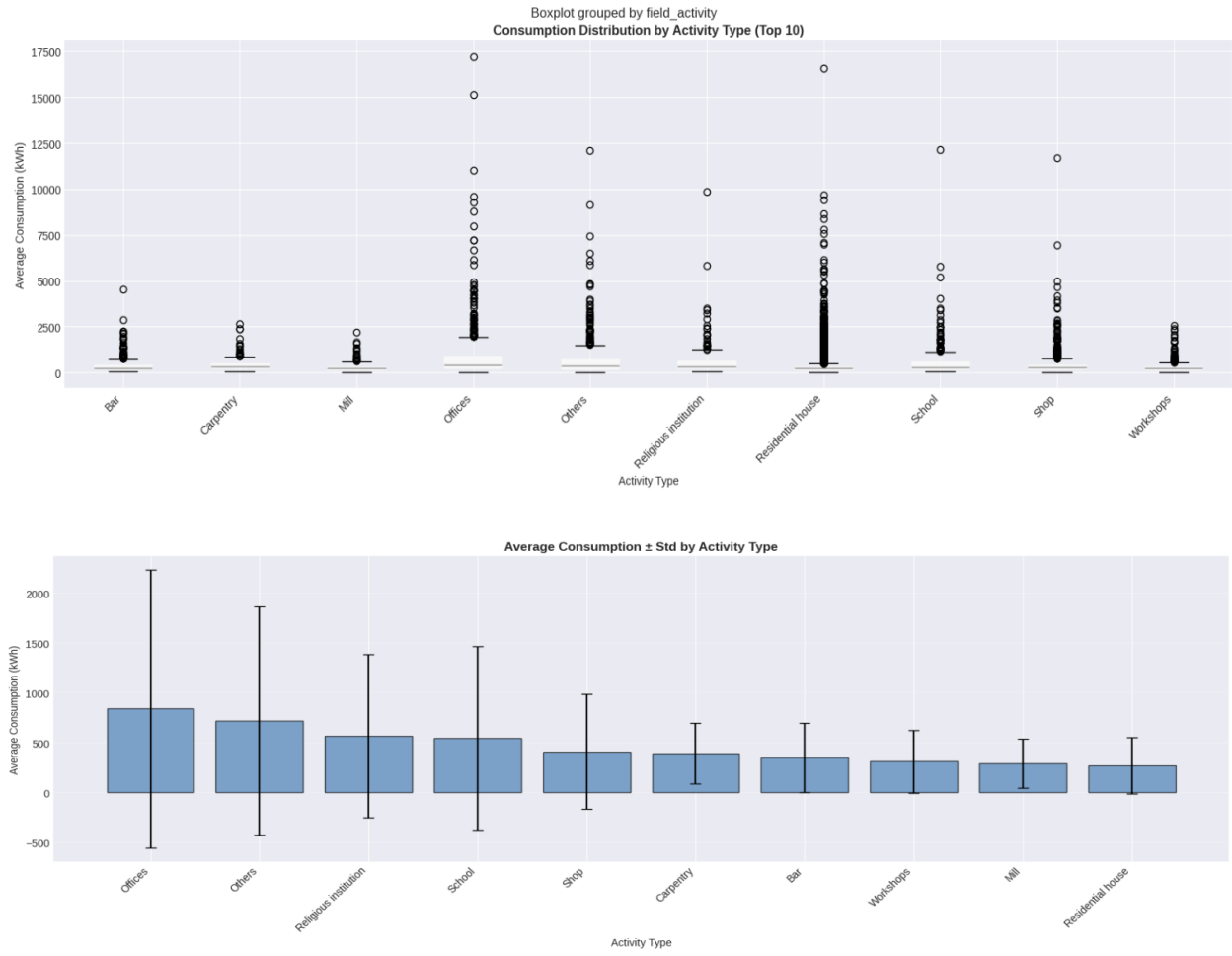


Fig. 10: Graph of activity analysis

DOI: 10.3844/ajeassp.2026.117.132

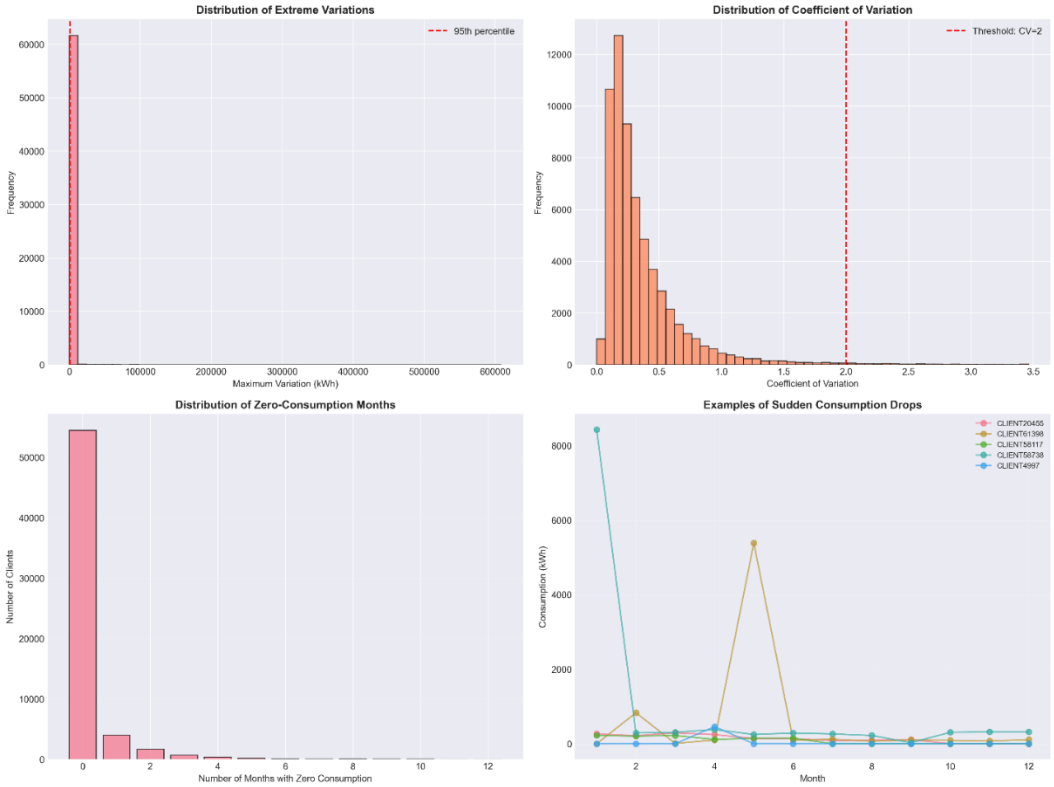


Fig. 11: Anomalies include extreme variations

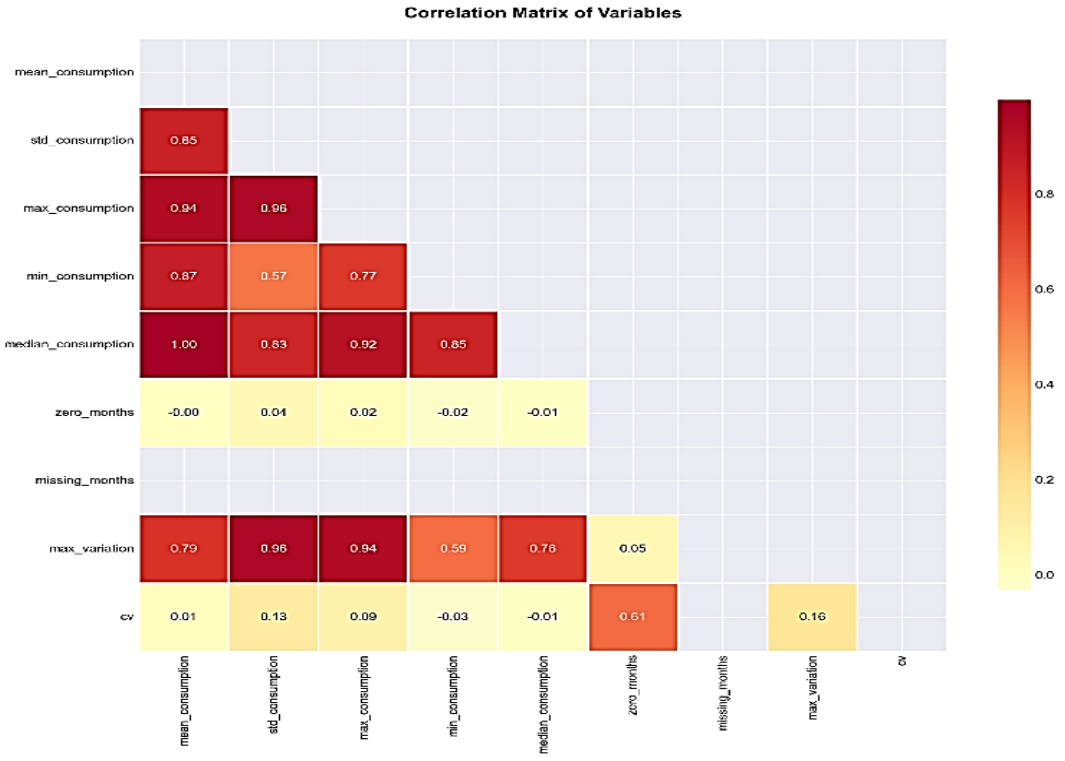


Fig. 12: Correlation matrix between the numerical variables

Statistical and Justified Determination of Thresholds

To transform the MSE error into a relevant risk classification (Suspicious or Fraudulent), thresholds were established statistically by analyzing percentiles, which limits false positives by targeting only the most extreme deviations in the error distribution. The adaptive thresholds chosen are:

- Suspicious threshold: Set at the 93rd percentile of the MSE error, i.e., 10.634911
- Fraudulent threshold: Set at the 97th percentile of the MSE error, i.e., 37.583787

Applying these thresholds enabled clear classification, identifying 93.0% of customers as Normal, 4.0% as Suspicious, and 3.0% as Fraudulent out of a total of 61,790 customers, confirming the clear separation of risk profiles.

Model Stability and Robustness

The stability of the model has been demonstrated through rigorous experimentation over different periods (training, validation, and test sets), ensuring that the results are not the result of overfitting or sensitivity to outliers:

- Overfitting Management: The model was optimized by increasing L2 regularization (to 0.01), increasing dropout (to 0.4), and simplifying the architecture (8 latent neurons). The validation/training performance ratio is 1.690, which is considered “slight overfitting” but remains acceptable
- Outlier Resistance: Using RobustScaler for normalization and clipping outliers at the 99th percentile ensures that the model has not been skewed by extreme data when learning normal profiles
- Built-in Anomaly Indicators: Reliability is enhanced by the inclusion of strong engineering features, such as missing value flags (e.g., high_missing_flag for ≥ 6 missing months) and extreme variation indicators, which act as behavioral evidence and reinforce the anomaly signal captured by the reconstruction error

All these optimizations confirm the relevance of the model as a fraud detection tool, as it has been designed to be both sensitive to anomalies and statistically robust.

Discussion

Model Performance

The results obtained show that the autoencoder performs better overall than the other artificial neural network architectures considered, particularly in terms of its ability to represent and detect the underlying structures of the data. Thanks to its unsupervised learning, the autoencoder is able to efficiently extract relevant features from raw data without requiring prior annotations, as presented in Table 2.

Comparison With Existing Literature

Compared to traditional supervised neural networks such as the Multilayer Perceptron (MLP), the autoencoder demonstrates a greater ability to model the normal behaviour of the system, resulting in lower reconstruction error for normal data and increased sensitivity to atypical observations. This property is particularly advantageous in contexts where labelled data is scarce or costly to obtain. Furthermore, the use of the latent space learned by the autoencoder allows for effective dimensionality reduction while preserving essential information. When this latent space is used as input for other ANNs (MLPs, classification networks), a notable improvement in performance is observed, particularly in terms of learning stability and reduced overfitting. These findings are consistent with the comprehensive review of deep learning presented by LeCun et al. (2015); Schmidhuber (2015), which highlight the superior capability of deep neural networks to extract hierarchical representations from complex datasets.

Practical Implications for Cameroon

The high predictive performance suggests that ANN-based detection systems can significantly reduce financial losses for the Cameroonian electricity distribution network. By automating fraud detection, utilities such as the National Electricity Company ENEO Cameroon can allocate field inspections more efficiently, targeting customers with the highest likelihood of irregularities. This reduces operational costs while improving revenue collection.

However, the implementation of such models must consider data privacy, cybersecurity, and adaptation to local consumption patterns, which may differ from patterns in other regions. Future research could incorporate temporal consumption trends, geospatial data, and smart meter readings to enhance model robustness.

Table 2: Numerical comparison with other methods

Method	Detection Accuracy	Sensitivity to Subtle Anomalies	Robustness to Outliers	Computational Complexity	Notes
Autoencoder (Deep Learning)	High (precision & recall often >85% in benchmark datasets)	Excellent – captures non-linear and temporal patterns	High – robust to extreme values due to reconstruction learning	High – requires GPU for large datasets and careful tuning	Performs best for complex, high-dimensional, or sequential data; reconstruction error provides an interpretable anomaly score
Rule-Based / Threshold Methods	Moderate – depends heavily on rules	Poor – cannot detect anomalies outside predefined thresholds	Low – extreme values may trigger false positives	Low – very fast and simple to implement	Works for obvious or known fraud patterns; fails on evolving or subtle anomalies.
Statistical Methods (z-score, Mahalanobis)	Moderate – effective if data follow assumed distributions	Moderate – may detect strong deviations but misses subtle non-linear patterns	Low to moderate – sensitive to outliers, unless robust statistics are used	Low-computationally efficient	Better for small datasets or when assumptions hold; less effective on high-dimensional or skewed data
Classical ML (Isolation Forest, One-Class SVM, kNN)	High in structured data (~75–85%)	Good – can detect non-linear outliers with appropriate features	Moderate – some sensitivity to extreme outliers	Medium – depends on dataset size and model complexity	Effective for unsupervised anomaly detection; performance improves with feature engineering, but may miss temporal trends

Conclusion

The application of neural networks in detecting and reducing non-technical losses within electrical distribution networks represents a significant advance in the field of smart energy management. By leveraging machine learning and big data analytics capabilities, these models can identify consumption irregularities, locate fraud, and anticipate losses with greater accuracy. This approach helps improve the operational efficiency of network managers while enhancing the transparency and reliability of the distribution system. In a context marked by growing energy demand and major economic and environmental challenges, the integration of artificial intelligence appears to be an essential strategic solution for optimizing the performance of electrical infrastructures. In the specific case of our study, this application of artificial intelligence provided us with the information we needed to take action for the customers of the National Electricity Company of Cameroon with a high degree of accuracy.

Acknowldgment

I am thankful to Ing Eyenga Minkonda Laurentine Séréna, who helped me with the implementation of the software.

Author's Contributions

Lekini Nkodo Claude Bernard: Contributed to research design, conceptualization, literature review, model development, and manuscript preparation and revision.

Bell Serge Samuel: Supported Claude day to day, helped with statistical analysis, and edited the manuscript.

Nyemb Nsoga Valjacques: Provided access to a dataset and some facilities, contributed to manuscript review.

Urbain Nzotcha: Contributed to manuscript review, revision, and research supervision.

Ethics

These authors have no ethical issues that may arise after the publication of this manuscript and confirm that this work is original and has not been published elsewhere.

References

Althobaiti, A., Jindal, A., Marnerides, A. K., & Roedig, U. (2021). A data-driven survey on fraud detection in smart grids: Techniques and challenges. *IEEE Access*, *9*.

Badr, M., Ibrahem, M., Kholidy, H., Fouda, M., & Ismail, M. (2023). Data-driven detection methods in smart metering systems: A comprehensive review. *Energies*, *16*(6), 2852.

Buzau, M. M., Tejedor-Aguilera, J., Cruz-Romero, P., & Gomez-Exposito, A. (2019). Detection of Non-Technical Losses Using Smart Meter Data and Supervised Learning. *IEEE Transactions on Smart Grid*, *10*(3), 2661–2670. https://doi.org/10.1109/tsg.2018.2807925

Carr, D., & Thomson, M. (2022). Scope review on non-technical losses in power systems. *Energies*, *15*(6), 2218.

Chollet, F. (2017). *Deep learning with Python*.

Depuru, S. S. S. R., Wang, L., & Devabhaktuni, V. (2011). Electricity theft: Overview, issues, prevention, and a smart meter-based approach to control theft. *Energy Policy*, *39*(2), 1007–1015. https://doi.org/10.1016/j.enpol.2010.11.037

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*.

Guarda, F., Hammerschmitt, B., Capeletti, M., Neto, N., dos Santos, L., Prade, L., & Abaide, A. (2023). Review of non-hardware-based detection of non-technical losses. *Energies*, *16*(4), 2054.

Haykin, S. (1999). *Neural networks: A comprehensive foundation*.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Kgaphola, P. M., Marebane, S. M., & Hans, R. T. (2024). Advanced methods for NTL detection in developing countries: Tools and challenges. *Energies*, *5*(2), 334–350.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791

Leite, J. B., & Mantovani, J. R. S. (2018). Non-technical loss detection and localization in modern distribution grids. *Electric Power Systems Research*, *9*(2), 1023–1032.

Lekini, N. C. B., Benoît, N., & Hamandjoda, O. (2017). Evaluation of Customer Behaviour Irregularities in Cameroon Electricity Network using Support Vector Machine. *American Journal of Engineering and Applied Sciences*, *10*(1), 32–42. https://doi.org/10.3844/ajeassp.2017.32.42

McCulloch, W. S., & Pitts, W. (1943). A logical Calculus of the Ideas Immanent in Nervous Activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133. https://doi.org/10.1007/bf02478259

Gong, X., Tang, B., Zhu, R., Liao, W., & Song, L. (2020). Data Augmentation for Electricity Theft Detection Using Conditional Variational Auto-Encoder. *Energies*, *13*(17), 4291. https://doi.org/10.3390/en13174291

Monedero, I., Biscarri, F., León, C., Guerrero, J. I., Biscarri, J., & Millán, R. (2006). Non-technical loss detection based on genetic algorithms and support vector machines. *Proceedings of the International Conference on Intelligent Systems Applications to Power Systems (ISAP)*. International Conference on Intelligent Systems Applications to Power Systems (ISAP), Abstract/Paper presented online / Distributed internationally via IEEE.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Saeed, M. S., Mustafa, M. W., Hamadneh, N. N., Alshammari, N. A., Sheikh, U. U., Jumani, T. A., Khalid, S. B. A., & Khan, I. (2020). A systematic review on non-technical loss detection methods. *IEEE Access*, *8*, 168916–168936.

Schmidhuber, J. (2015). Deep Learning in Neural Networks: An Overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Stracqualursi, E., Rosato, A., Lorenzo, G. D., Panella, M., & Araneo, R. (2023). Socio-economic analysis in NTL detection: AI and policy implications. *Electric Power Systems Research*, *184*, 113544.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Proceedings of the Advances in Neural Information Processing Systems*, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762

Viegas, J. L., Esteves, P. R., Melício, R., Mendes, V. M. F., & Vieira, S. M. (2017). Data analytics for fraud detection: A critical review. *Electric Power Systems Research*, *147*, 196–209.

Xia, X., Xiao, Y., Liang, W., & Cui, J. (2022). Review of electricity theft detection methods using smart meters. *IEEE Transactions on Smart Grid*, *13*(2), 1245–1258.