Gazi Üniversitesi
**Fen Bilimleri Dergisi**
PART C: TASARIM VE TEKNOLOJİ

Gazi University
**Journal of Science**
PART C: DESIGN AND TECHNOLOGY

GU J Sci, Part C, 12(2): 438-456 (2024)

# Electricity Theft Detection Using Rule-Based Machine Learning (rML) Approach

Sheyda BAHRAMI[1] Erol YUMUK[1] Alper KEREM[1,2*] Beytullah TOPCU[1] Ahmetcan KAYA[1]

[1] *NAR System Technology Inc (NAR Sistem Teknoloji A.Ş.), Istanbul, Turkey*

[2] *Kahramanmaras Sutcu Imam University, Faculty of Engineering and Architecture, Department of Electrical and Electronics Engineering, Kahramanmaras, Turkey*

### *Article Info*

*Research article*
*Received: 27/02/2024*
*Revision: 12/03/2024*
*Accepted: 18/03/2024*

### *Keywords*

*Electricity Theft Detection*
*Non-Thechnical Losses*
*Advanced Metering Infrastructure*
*Machine Learning*

### *Makale Bilgisi*

*Araştırma makalesi*
*Başvuru: 27/02/2024*
*Düzeltme: 12/03/2024*
*Kabul: 18/03/2024*

### *Anahtar Kelimeler*

*Kaçak Elektrik Tespiti*
*Teknik Olmayan Kayıplar*
*İleri Ölçüm Altyapı*
*Makine Öğrenimi*

### *Graphical/Tabular Abstract (Grafik Özet)*

*In this study, various machine learning techniques combined with a new rule-based feature space were utilized to identify electricity theft. / Bu araştırmada, yeni bir kural tabanlı özellik uzayıyla birleştirilen çeşitli makine öğrenimi yöntemleri elektrik kaçağını belirlemek amacıyla kullanılmıştır.*

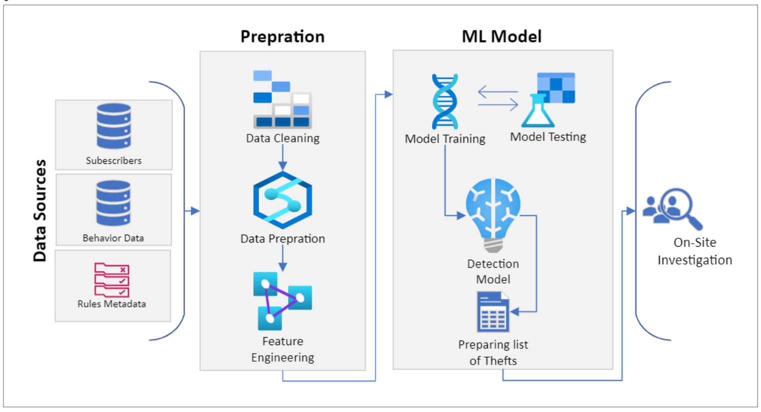



***Figure A:*** *The block diagram of proposed method /* ***Şekil A:*** *Önerilen yöntemin blok diyagramı*

### *Highlights (Önemli noktalar)*

- *Development of a unique hybrid approach that integrates expert-derived rules with data-driven models. / Veri modelleri ile uzman kurallarının birleştirilmesinden oluşan özgün bir yaklaşım geliştirilmesi.*
- *Incorporation of nine specific expert-investigated rules into the models. / Modellere dokuz özel uzman kuralının eklenmesi.*
- *Implementation of on-site evaluations to validate the model's predictions. / Modellerin tahminlerini test etmek için saha incelemeleri yapılması.*
- *Comprehensive evaluation of load histories from both fraudulent and normal cases. / Kaçak ve normal vakalardan gelen yük geçmişlerinin kapsamlı değerlendirilmesi.*
- *Exploration and comparison of eight supervised machine learning models. / Sekiz denetimli makine öğrenimi modelinin incelenmesi ve karşılaştırılması.*

***Aim (Amaç):*** *The aim of this article is to introduce a novel, rule-based combined machine learning technique for detecting electricity theft. / Bu makalenin amacı, elektrik kaçak tüketimini tespit etmek için yeni bir kural tabanlı makine öğrenimi tekniği tanıtmaktır.*

***Originality (Özgünlük):*** *In this study, unlike traditional approaches that focus solely on consumption patterns, our methodology integrates unique, expert-derived insights, significantly advancing the effectiveness of fraud detection techniques. / Bu çalışma, geleneksel tüketim odaklı yaklaşımların aksine uzman içgörülerini entegre ederek kaçak tespitinde önemli bir yenilik sunmaktadır.*

***Results (Bulgular):*** *Ensemble Methods, especially Random Forest, dominated in recall performance (0.93), while AdaBoost, LGBoost and XGBoost also showed strong results. / Topluluk Yöntemleri, özellikle de Rastgele Orman, hatırlama performansında (0,93) üstünlük sağlarken AdaBoost, LGBoost ve XGBoost da güçlü sonuçlar göstermiştir.*

***Conclusion (Sonuç):*** *Random Forest stood out for its ability to handle diverse data types and minimize overfitting, accurately identifying 77% of theft instances in on-site inspections. / Rastgele Orman, farklı veri türleriyle başa çıkma ve aşırı uyumu en aza indirme becerisiyle öne çıkarak yerinde denetimlerde hırsızlık örneklerinin %77'sini doğru bir şekilde tespit etmiştir.*

*Corresponding author, e-mail: alperkerem@ksu.edu.tr
DOI: 10.29109/gujsc.1443371

GU J Sci, Part C, 12(2): 438-456 (2024)

Gazi Üniversitesi
**Fen Bilimleri Dergisi**
PART C: TASARIM VE TEKNOLOJİ

Gazi University
**Journal of Science**
PART C: DESIGN AND TECHNOLOGY

http://dergipark.gov.tr/gujsc

# Electricity Theft Detection Using Rule-Based Machine Learning (rML) Approach

Sheyda BAHRAMI[1] Erol YUMUK1 Alper KEREM[1,2*] Beytullah TOPCU[1] Ahmetcan KAYA[1]

[1] *NAR System Technology Inc (NAR Sistem Teknoloji A.Ş.), Istanbul, Turkey*

[2] *Kahramanmaras Sutcu Imam University, Faculty of Engineering and Architecture, Department of Electrical and Electronics Engineering, Kahramanmaras, Turkey*

***Article Info***

*Research article*
*Received: 27/02/2024*
*Revision: 12/03/2024*
*Accepted: 18/03/2024*

***Keywords***

*Electricity Theft Detection*
*Non-Thechnical Losses*
*Advanced Metering Infrastructure*
*Machine Learning*

**Abstract**

Since electricity theft affects non-technical losses (NTLs) in power distribution systems, power companies are genuinely quite concerned about it. Power companies can use the information gathered by Advanced Metering Infrastructure (AMI) to create data-driven, machine learning-based approaches for Electricity Theft Detection (ETD) in order to solve this problem. The majority of data-driven methods for detecting power theft do take usage trends into account while doing their analyses. Even though consumption-based models have been applied extensively to the detection of power theft, it can be difficult to reliably identify theft instances based only on patterns of usage. In this paper, a novel rule-based combined machine learning (rML) technique is developed for power theft detection to address the drawbacks of systems that rely just on consumption patterns. This approach makes use of the load profiles of energy users to establish rules, identify the rule or rules that apply to certain situations, and classify the cases as either legitimate or fraudulent. The UEDAS smart business power consumption dataset's real-world data is used to assess the performance of the suggested technique. Our technique is an innovation in theft detection that combines years of intensive theft tracking with the use of rule-based systems as feature spaces for traditional machine learning models. With an astounding 93% recall rate for the rule-based feature space combination of the random forest classifier, this novel approach has produced outstanding results. The acquired results show a noteworthy accomplishment in the field of fraud detection, successfully detecting fraudulent consumers 77% of the time during on-site examination.

# Kural Tabanlı Makine Öğrenimi (rML) Yaklaşımı Kullanılarak Kaçak Elektrik Tespiti

***Makale Bilgisi***

*Araştırma makalesi*
*Başvuru: 27/02/2024*
*Düzeltme: 12/03/2024*
*Kabul: 18/03/2024*

***Anahtar Kelimeler***

*Kaçak Elektrik Tespiti*
*Teknik Olmayan Kayıplar*
*İleri Ölçüm Altyapı*
*Makine Öğrenimi*

**Öz**

Elektrik kaçağının güç dağıtım sistemlerinde teknik olmayan kayıplar (NTL'ler) üzerindeki etkisi göz önüne alındığında, enerji şirketleri bu soruna büyük bir önem atfetmektedir. Enerji şirketleri, bu problemi çözmek amacıyla, Kaçak Elektrik Tespiti (ETD) için İleri Ölçüm Altyapısı (AMI) tarafından toplanan verileri kullanarak veriye dayalı, makine öğrenimine dayanan yöntemler geliştirebilir. Elektrik kaçağını tespit etmeye yönelik mevcut veriye dayalı metodolojiler, analizlerini gerçekleştirirken genellikle kullanım trendlerini hesaba katmaktadır. Tüketim bazlı modellerin elektrik kaçağının tespitinde yaygın olarak uygulanmış olmasına rağmen, yalnızca kullanım desenlerine dayanarak kaçak vakalarını güvenilir bir şekilde tanımlamak zorluklar içerebilir. Bu çalışmada, tüketim desenlerine dayalı sistemlerin kısıtlılıklarını ele almak üzere, enerji kullanıcılarının yük profillerini kullanarak kurallar oluşturmak, belirli durumlar için uygulanabilir kural veya kuralları belirlemek ve vakaları normal veya kaçak olarak sınıflandırmak üzere kural tabanlı birleşik bir makine öğrenimi (rML) tekniği geliştirilmiştir. UEDAŞ akıllı iş gücü tüketim veri setinin gerçek dünya verileri, önerilen yöntemin performansının değerlendirilmesinde kullanılmıştır. Bu yöntem, geleneksel makine öğrenimi modelleri için özellik uzayları olarak kural tabanlı sistemlerin kullanılmasıyla yıllar süren yoğun hırsızlık takibinin entegrasyonunu temsil eden kaçak tespitinde bir yeniliktir. Kural tabanlı özellik uzayının rastgele orman sınıflandırıcısı ile kombinasyonu için %93 gibi dikkat çekici duyarlılık oranı ile bu yeni yaklaşım, saha incelemeleri sırasında kaçak faaliyetlerini %77 oranında başarıyla tespit ederek olağanüstü sonuçlar üretmiştir.

*Corresponding author, e-mail: alperkerem@ksu.edu.tr
DOI: 10.29109/gujsc.1443371

## 1. INTRODUCTION (GİRİŞ)

Machine learning offers promising solutions for anti-electricity stealing by analyzing users' electricity consumption behavior. However, in the real world, due to different patterns of electricity consumption and imbalanced numbers of fraudulent activities, these solutions may fail to generalize effectively. Electricity theft is a significant problem that affects the normal operation of power grids and causes economic losses for power enterprises. In developing countries, the electricity theft rate can be as high as 30%, resulting in substantial financial impacts on power suppliers [1]. According to recent research, the global annual loss due to electricity theft amounts to an astonishing $96 billion [2]. To combat this issue, efficient anti-electricity theft measures must be implemented to ensure a reasonable power supply and rational use of electricity, thereby reducing economic losses as much as possible [1]. Electricity theft detection methods can be broadly categorized into three categories: manually-driven, physically-driven, and data-driven. The manually-driven method relies on technicians manually checking electricity meters one by one, which is time-consuming and labor-intensive. The physically-driven method analyzes physical rules or sensor data in the grid to detect power theft, but it can be costly and requires specific sensors and topology information. The implementation of advanced metering infrastructure (AMI) in smart grids has led to an increase in the use of data-driven methods, which are now more prevalent. AMI enables the collection of large amounts of electrical consumption data at high frequencies, making it useful for electricity theft detection. Data-driven analysis algorithms can detect anomalies in the data and play a crucial role in identifying theft at a lower cost [2]. These methods can be categorized into unsupervised learning, semi-supervised learning, and supervised learning, depending on the level of prior knowledge required. Unsupervised learning does not require labeled data, semi-supervised learning uses a small amount of labeled data, and supervised learning relies on a significant amount of labeled data for detection [3]. Several research papers discuss methods and approaches for detecting electricity theft using machine learning and data analysis techniques. These studies focus on analyzing users' electricity consumption behavior and identifying anomalies that may indicate fraudulent activities. By detecting abnormal consumption patterns, it is possible to identify potential instances of theft and improve the efficiency of anti-stealing efforts. Deep learning approaches, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and generative adversarial networks (GANs), have also shown promise in detecting abnormal electricity consumption patterns and identifying theft [4]. Supervised learning methods in electricity theft detection such as Support Vector Machines (SVM), decision trees, and ensemble learning methods like XG-Boost, have been employed in this context with differing levels of success [2]. These methods utilize Artificial Intelligence (AI) techniques to analyze energy consumption data and identify anomalies that may indicate theft or other irregularities. By leveraging large amounts of data from sub-meters and smart sensors, it becomes possible to detect anomalous power consumption and understand the causes of each anomaly [5]. The choice of detection method depends on factors such as the availability of labeled data, data quality, complexity of the power grid environment, and desired detection accuracy. Researchers continue to explore and develop more sophisticated and efficient methods for electricity theft detection to overcome the limitations of existing approaches [2]. Feature engineering and structured query language (SQL) analytic functions have also been proposed as solutions for detecting electricity frauds using machine learning. These approaches aim to engineer relevant features from the data and utilize SQL analytic functions to detect fraudsters. By improving the correlation between data features and the target variable, these methods enhance the model's ability to identify fraudulent activities [6]. Table 1 presents the data-driven work conducted on the detection of electricity theft. Although machine learning approaches hold promise for electricity theft detection, the effectiveness of these methods may vary depending on the specific patterns of electricity consumption and the distribution of fraudulent activities in different regions and contexts.

**Table 1**. Literature on electricity theft detection (ETD) studies (Elektrik hırsızlığı tespiti (ETD) çalışmaları üzerine literatür)

| Consumer type | Definition | Models | Recall | F1-Score | Precision | Reference |
|---|---|---|---|---|---|---|
| None-public | high-contracted power consumers in Endesa Company, Spain, analyzing their two-year consumption data when the consumption exceeds 1000 kW. | Bayesian networks, decision trees and Pearson correlation | 0.38 | - | - | [7] |
| Feeder Data | The data, spanning from February 1st, 2017 to April 30th, 2019, encompass parameters of power equipment, load, electricity sales, and records of electricity ammeter openings at the Tianjin Electric Power Company in China | Support vector machines | - | - | - | [8] |
| Industrial and commercial | Dataset spans 22 months, covering the period from May 2014 to February 2016 by the Company of Electric Energy of Honduras | Support vector machines | 0.33 | 0.18 | 0.13 | [9] |
| None-public | consumption data from Ireland, solar generation data from the U.S. and Australia, and wind generation data from France | PCA, kullback-Leibler divergence, density-based clustering | 0.87 | - | - | [10] |
| Residence | Using the demand data from more than 4000 households over an 18-months period | Finite mixture model clustering, GBM | - | - | - | [11] |
| None-public | Data are performed using State Grid Corporation of China (SGCC) dataset | K-means, local outlier factor | - | - | - | [12] |
| Residence | The dataset encompasses a variety of measurements from multiple sensors, capturing energy usage, occupancy, and ambient conditions within a household. The data were collected over a six-month period, spanning from July 5th to December 5th, 2015 | K-Means clustering | - | - | - | [13] |
| Residence | The Dutch Residential Energy Dataset (DRED), specifically its public residential dataset, is utilized in this study. This dataset encompasses a variety of measurements from multiple sensors, capturing energy usage, occupancy, and ambient | Fuzzy C-means clustering | 0.17 | - | 0.83 | [14] |

| | | | | | | |
|---|---|---|---|---|---|---|
| | conditions within a household. The data were collected over a six-month period, spanning from July 5th to December 5th, 2015 | | | | | |
| Residence | Data of approximately 5000 Irish households monitored for one and a half years | ANNs and Fuzzy set theory (ANFIS classification method) | 0.99 | 0.99 | 0.99 | [15] |
| Industrial and commercial | Two private datasets from a Brazilian electric utility, represented by eight features (Demand Billed, Demand Contracted, Maximum Demand, Reactive Energy, Power Transformer, Power Factor, Installed Power, Load Factor) | Probabilistic OPF | - | - | - | [16,17,18,19] |
| Industrial and commercial | Two private datasets from a Brazilian electric utility, represented by eight features (Demand Billed, Demand Contracted, Maximum Demand, Reactive Energy, Power Transformer, Power Factor, Installed Power, Load Factor) | Genetic algorithms, harmony search, OPF, particle swarm optimization | - | - | - | [20] |
| IEEE 34-bus test case | Electricity consumption dataset | Random matrix theory | 0.91 | - | - | [21] |
| None-public | Electricity consumption dataset | Rule engine, SVM | - | - | - | [22] |
| Farmers and commercial | Smart meter data collected from 171 consumers at a 15-minute resolution in Nana Kajaliyala village, Gujarat, India. | Hierarchical clustering and decision tree | - | - | - | [23] |
| Residence | Master meter and smart meter dataset (114 single family apartment during one year) in Western Massachusetts | Correlation analysis, Pearson correlation | - | 0.8 | - | [24] |
| Residence | In the simulations, the load profiles are derived from actual residential active power consumption data. This data is recorded at a granularity of one minute, representing the average power usage within each minute. | - | - | - | - | [25] |

| | | | | | | |
|---|---|---|---|---|---|---|
| Hard-ware | Numerical experiments on the Irish smart meter dataset are conducted to show the good performance of the combined method | Correlation analysis | | | | [26] |
| None-public | Electricity consumption data was built at University of Michigan | Covariance matrix | - | - | - | [27] |
| Residence | The historical electricity consumption sequence of users and the relevant information of similar users are obtained from electricity consumption database | DBSCAN | - | - | - | [28] |
| None-public | Real data of 12752 consumers is used from the power utilities in Pakistan on monthly basis | Multivariate Gaussian Distribution | 0.75 | - | - | [29] |
| Residence, Commercial | Theft dataset is injected to the consumption patterns | Entropy analysis | - | - | - | [30,31] |
| Residence | Real-time electricity theft detection using simulated resident energy consumption data | Special ETPS unit | - | - | - | [32] |
| Residence | Real-time electricity theft detection using energy consumption data of 5 residents | Prediction-based regression, prediction-based neural network, clustered-based, and projection-based methods | - | - | - | [33] |
| Residence | Dataset from a US electric utility, represented by five features (Load rate, minimum load coefficient, load rate during peak load period, load rate during stable load period, load rate during valley load period) | Local matrix construction | - | - | - | [34] |
| Residence | Massive SM data includes voltage, current, active power | Deep learning-based Semi-Supervised Auto-Encoder | 0.80 | 0.86 | 0.94 | [35] |
| None-public | Real-world-data-based case studies are presented, which have shown that adding unlabeled samples into training set has greatly improved the performance | Utilizing unlabeled data to detect electricity fraud in AMI: A semi-supervised deep learning approach | - | - | - | [36] |
| None-public | simulations are performed using State Grid | Self-Attention Generative Adversarial Network | 0.99 | 0.9 | 0.95 | [37] |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Corporation of China (SGCC) dataset | (SAGAN) is used in combination with Convolutional Neural Network (CNN) | | | | |
| IEEE 123-bus case | The functions of a cyber-attack are implemented on a benign dataset spanning one year to generate a malicious dataset. | CNN, GRU-RNN | 0.993 | 0.995 | 0.997 | [38] |
| Residence | Electricity consumption patterns of 90 low-voltage distribution customers over a three-month period from January 1st to March 31st, 2013, using a 30-minute resolution on the Low Carbon London smart meter trials dataset | K-means and DWT | - | - | - | [39] |
| None-public | Imbalanced realistic dataset that presents a daily electricity consumption provided by State Grid Corporation of China | CNN | - | - | - | [40] |
| Residence, Commercial | Smart meter data set provided by the Greek DSO, HEDNO and a publicly available smart meter data set. Frauds are simulated and the Twitter breakout detection library is used for extracting features | Rule systems, multi-variate Gaussian distribution (MGD), local outlier factor (LOF), k-means, fuzzy c-means, DBSCAN and SOM | 0.76 | 0.755 | - | [41] |
| None-public | State Grid Corporation of China (SGCC) dataset | VGG neural network, XGBoost | 0.97 | 0.937 | 0.93 | [42] |

Achieving generalization across diverse scenarios remains a challenge. Further research and development are necessary to improve the robustness and accuracy of these models in real-world applications. In this paper, we propose a method that surpasses AI methods solely reliant on consumption patterns. Our approach aims to identify fraudulent consumers by leveraging rules developed in collaboration with experts. These rules are extracted from the load profiles of fraudulent consumers. It is worth emphasizing that our method incorporates consumption-based rules, which constitute just one part of the 14 rules under consideration. Our approach in detecting electricity theft represents a significant advancement over many existing studies, which often emphasize data-driven methodologies based on analyzing consumption patterns. A notable limitation in these studies, is their reliance on datasets that simulate electricity theft scenarios rather than reflecting actual incidents. These simulated datasets may not accurately capture the real-world nuances of electricity theft. Moreover, a critical aspect often overlooked in data-driven methodologies is the lack of on-site investigation. By relying solely on theoretical or simulated data, these models miss the opportunity to incorporate crucial empirical insights that can only be derived from physical verification and real-world observations. This gap results in a significant disconnect between the model outputs and the actual on-ground scenarios of electricity theft, limiting the practical applicability and effectiveness of these models in real-life situations. In contrast, our model distinguishes itself by integrating 9 specific rules, thoroughly investigated by experts, as key features. This inclusion of expert insights allows for a more nuanced and informed analysis of consumption data, directly targeting the complexities of real-world theft patterns. Moreover, our model extends beyond mere data analytics. Following the predictive analysis, we conduct on-site evaluations to verify the model’s results. This dual approach of integrating expert-derived rules and conducting physical verifications on-site ensures a comprehensive and accurate detection of electricity theft. To effectively determine the rules that fraudulent consumers adhere to, we evaluate the load histories of fraudulent and normal cases. Subsequently, these rules are treated as feature spaces, enabling us to classify consumers as either theft or benign. The trained model is then deployed for real-world electricity theft detection. Throughout this study, we explore 8 supervised models to identify the most optimal one. Among these models, the random forest algorithm exhibits the highest performance, delivering accurate results.

The key contributions of this study are as follows:

- Development of a unique hybrid approach that integrates expert-derived rules with data-driven models for electricity theft detection.
- Incorporation of nine specific expert-investigated rules into the model, enhancing its accuracy and applicability to real-world scenarios.
- Implementation of on-site evaluations to validate the model's predictions, bridging the gap between theoretical analysis and practical fieldwork.
- Comprehensive evaluation of load histories from both fraudulent and normal cases, providing a more nuanced understanding of electricity theft patterns.
- Exploration and comparison of eight supervised machine learning models to determine the most effective algorithm for this application.
- Demonstration of the superiority of the random forest algorithm in our context, backed by empirical evidence from real-world data.

The rest of the articles are as follows. Section 2 describes the models. The method and analysis of experiment results is shown in Section 3. Finally, results and concluding remarks are presented in Section 4 and Section 5, respectively.

## 2. AI MODELS IN NON-TECHNICAL LOSS DETECTION (TEKNİK OLMAYAN KAYIP TESPİTİNDE YAPAY ZEKA MODELLERİ)

We examined a range of machine learning algorithms for NTL (Non-Technical Loss) detection, including Light-GBM, XG-Boost, Random Forest, Support Vector Machine (SVM), AdaBoost, K-Nearest Neighbors, Decision Trees and Multi-Layer Perceptron. Our goal was to identify the most effective classifiers for NTL detection.

### 2.1 ENSEMBLE METHODS (TOPLULUK YÖNTEMLERİ)

Ensemble methods in machine learning involve combining predictions from multiple base models to enhance overall performance. There are two main categories of ensemble methods:

a. Averaging-Based Ensembles: These methods aggregate results from individual models by averaging their predictions, resulting in superior performance compared to single models. Random forest [43] is an example of this category, which combines predictions from randomized decision trees.

b. Boosting-Based Ensembles: These techniques combine weak learners to create a robust ensemble, reducing prediction bias. Examples include AdaBoost [44], Light-GBM [45], and XG-Boost. AdaBoost assigns greater weights to incorrectly predicted instances, guiding the model towards better performance. Light-GBM uses a depth-first approach for quicker training but may overfit on smaller datasets. XG-Boost [45] requires preprocessing for categorical features and handles missing data.

Ensemble methods improve predictive performance by combining outputs from multiple models. Each method has its strengths and considerations, making them suitable for different machine learning use cases.

### 2.2 SUPPORT VECTOR MACHINE (DESTEK VEKTÖR MAKİNESİ)

Support Vector Machines (SVM) [46] are a versatile class of machine learning techniques used for tasks such as outlier detection, regression, and classification. They are widely adopted in data mining due to their strong predictive capabilities and reliability in supervised learning. In our data classification, we used the Linear Support Vector Classifier (Linear SVC).

### 2.3 DECISION TREES (KARAR AĞAÇLARI)

Decision Trees are a fundamental machine learning technique used for both classification and regression tasks. They work by creating a tree-like structure where each node represents a decision based on specific features of the data. Decision Trees are easy to understand and interpret, require minimal data preprocessing (no need for data normalization), and can handle both numerical and categorical data. They can create overly complex trees prone to overfitting. Ensuring optimal tree structure can be computationally intensive, and they may not perform well on very small datasets [23].

### 2.4 NEURAL NETWORK MODEL (SİNİR AĞI MODELİ)

A multi-layer perceptron (MLP) is a type of artificial neural network designed for various machine learning tasks, particularly in deep learning. It consists of multiple interconnected layers of artificial neurons, each layer playing a unique role in information processing. Typically, an MLP includes an input layer, one or more hidden layers, and an output layer. The input layer receives the initial data, which is then passed through the hidden layers. Neurons in these hidden layers apply mathematical transformations to the input data, learning and extracting complex patterns and features. The final output layer produces the network's prediction or classification. MLPs are known for their ability to model complex relationships in data and are used in tasks such as image recognition, natural language processing, and predictive modeling. They are a fundamental component of deep learning, contributing to the success of modern artificial intelligence applications [47].

## 3. METHODOLOGY (METODOLOJİ)

### 3.1. FEATURE EXTRACTION (ÖZELLİK ÇIKARIMI)

To effectively combat electricity theft, it is necessary to extract comprehensive features from abnormal electricity consumption phenomena and quantifiable characteristics caused by various electricity theft behaviors.

Our team of experts has conducted a comprehensive and methodical examination of electricity theft patterns. Through this diligent analysis, we have formulated the following 14 rules to accurately assess the characteristics of a user's electricity consumption:

1. Meter tampering warning: Someone might have physically tampered with the smart meter to modify its internal components, leading to inaccurate readings. In these cases, the warning of the body lid being opened is checked. For suspicion of fraud, it should be evaluated along with sudden decrease in consumption.

2. Virtual meter Control: The accuracy of the measurements made by the meter is regularly verified. Every 15 minutes, the product of the current, voltage, and power factor (cosine of the angle between current and voltage) is calculated and compared with the instantaneous value measured by the meter. Due to the nature of the meter, which provides instantaneous readings, and the control being performed every 15 minutes, there may be deviations with a tolerance of 10%. These deviations are considered acceptable and are not flagged as out of control. The active and reactive power progress, calculated based on the current, voltage, and power factor, is continuously monitored. If the calculated value exceeds the measurement value by 10%, it is flagged for further investigation. This step is taken to ensure that any significant discrepancies are promptly identified and addressed.

3. Daily average voltage V: In the context of normal users, the voltage typically exhibits minimal fluctuations, indicating a relatively stable power consumption pattern. However, if there are noticeable abnormalities or deviations from the expected voltage levels, it could be an indication of irregular power consumption behavior by the user. These fluctuations may arise from various factors, such as faulty electrical equipment, overloading of circuits, or improper power usage. As such, monitoring the daily average voltage becomes crucial in identifying potential issues and ensuring the efficient and reliable operation of the electrical system.

4. Excessive number of frozen electricity amounts: If one of the voltage phases is continuously removed, the number of interruptions will increase. Increasing the number of interruptions will give a clue that one of the voltage phases has been removed repeatedly. The case of more than 500 interruptions in a month is checked. If the number of interruptions exceeds 500 times in a month, it should be considered as a suspicious

indicator of potential electricity theft. Frequent and excessive interruptions in the power supply, especially in such high numbers, could signal irregularities in electricity consumption.

5. General Current Control: In a typical electrical system, the metering current exhibits irregular fluctuations depending on the user’s load access. However, under normal operating conditions, the phase line current and neutral line current for a user should be nearly equal. This means that the total current flowing through the user’s electrical circuit should be balanced, with minimal deviation between the current in the phase line and the neutral line. By closely monitoring and comparing the phase line and neutral line currents, it becomes possible to detect any discrepancies or imbalances in the system. Significant differences between these currents may indicate electrical faults, leakage, or other issues that warrant attention.

6. Installed Power-Demand Control: In a well-functioning electrical system, the installed power should not surpass the demand (maximum consumption during a month). Any instance where the demand exceeds the installed power requires immediate attention. Specifically, the system continuously monitors the power demand, and in cases where it exceeds the installed capacity or reaches more than 20% beyond the contracted limit for the day, specific measures are taken to rectify the situation

7. Demand Correlation Evaluation: The demand correlation analysis is performed by comparing data from corresponding months over a span of two years. In this study, the Grey Correlation Grade (GCG) [48] method is employed to calculate the correlation coefficient. This approach helps us identify and assess the level of correlation between the demand patterns of different subscribers during similar months over the two-year period. If the resulting correlation coefficient is found to be below 0.6, the subscribes data is classified as a potential candidate for the illegal suspect list. Given two data series $X_0$ and $X_i$, the GCG can be calculated;

$$X_0^*(k) = \frac{X_0(k) - \min(X_0(k))}{\max(X_0(k) - \min(X_0(k))} \text{ for k =1,2, ..., n} \quad (1)$$

$$X_i^*(k) = \frac{X_i(k) - \min(X_i(k))}{\max(X_i(k) - \min(X_i(k))} \text{ for k=1,2, ..., n} \quad (2)$$

where min $(X_0(k)$ and min $(X_i(k)$represent the minimum values in $X_0$ and $X_i$, and max $(X_i(k)$ and max $(X_i(k)$ represent the maximum values in $X_0$ and $X_i$, respectively.

$$\zeta(k) = \frac{\xi \max(\Delta) + \min(\Delta)}{\Delta_{oi}(k) + \xi\, \mathrm{mac}(\Delta)} \text{ for k=1,2,…, n} \quad (3)$$

where $\xi$ is the distinguishing coefficient, set to 0.5 as per reference [8], $\Delta = |X_0^*(k) - X_i^*(k)|$, and $\Delta_{oi}(k)$ = $|X_0(k)$ -$X_i(k)|$.

$$\mathrm{GCG} = \frac{1}{n}\sum_{k=1}^{n} \zeta(k) \quad (4)$$

8. Demand Consumption Control (Shift Control): If the operating hours calculated based on demand and consumption data result in 3 hours for a facility that normally consumes 50,000 kWh by operating 8 hours a day, the reason for this decrease in scale should be investigated. The working hours of the sectors are determined by experts.

9. Current-demand Control: It is not possible that a meter records current without demand. During the first reading of the month, there might be instances where a "current available, no demand" alarm is received due to the fact that demand data has not been formed yet. In cases where there is a current of 0.1 A or more in any phase, the meter will be checked to ensure that demand recording is functioning correctly.

10. Consumption Correlation in Two Years: Comparing consumption over two years helps see if a customer's consumption patterns changed or not. This correlation is quantified as a number between -1 and 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation.

11. Eight Month Consumption Correlation: For each consumer, the correlation between the last eight months of electricity consumption across two consecutive years has been calculated. The values of this correlation are expressed on a scale similar to the one outlined in item 10, where the strength of the relationship between the consumption patterns is quantified numerically.

12. Daily Average Consumption: The average consumption of consumers during peak hours is calculated and then compared with their

installed power capacity. If this average consumption falls below 33% of the installed power, the consumer is flagged as potentially suspicious.

13. Three Month Consumption Correlation: As described in item 10, this involves a calculation of the correlation of electricity consumption. However, instead of spanning eight months, this assessment focuses on a three-month period.

14. Daily Average Demand: The daily average demand for electricity is calculated for two consecutive years and compared. If the value from the most recent year is at least 33% lower than that of the previous year, this may indicate that the consumer's usage is malignant.

### 3.2 ELECTRICITY THEFT DETECTION BASED ON EXTRACTED FEATURES

(ÇIKARILAN ÖZELLİKLERE DAYALI ELEKTRİK HIRSIZLIĞI TESPİTİ)

With the guidance and expertise of industry specialists 14 rules were instantiated within a Java-based analytical framework. Over an observation period, we deployed these rules against our dataset to scrutinize the inter-correlations present among them. Notably, a subset of these rules exhibited a high degree of correlation, prompting an analysis-driven decision to excise the redundant rules from our framework. Consequently, we distilled our feature space down to 9 core rules. Despite the high correlation observed among certain rules, expert consultation reinforced the decision to preserve them. This strategic choice was underpinned by the nuanced domain knowledge of our experts, ensuring that the integrity and depth of our analytical capabilities remain robust. Subsequent to the refinement of our feature space, the 9 retained rules were employed as inputs for our classification model. This approach was predicated on the hypothesis that a more streamlined and pertinent set of features would enhance the model's predictive performance. By leveraging a targeted feature set, informed by empirical evidence and domain expertise, we aimed to strike a balance between model complexity and classification accuracy. The results of this methodology are anticipated to validate the efficacy of our feature selection process in the realm of data-driven predictive modeling. Figure 1 illustrates the flowchart, and Figure 2 presents the block diagram of our method, respectively.

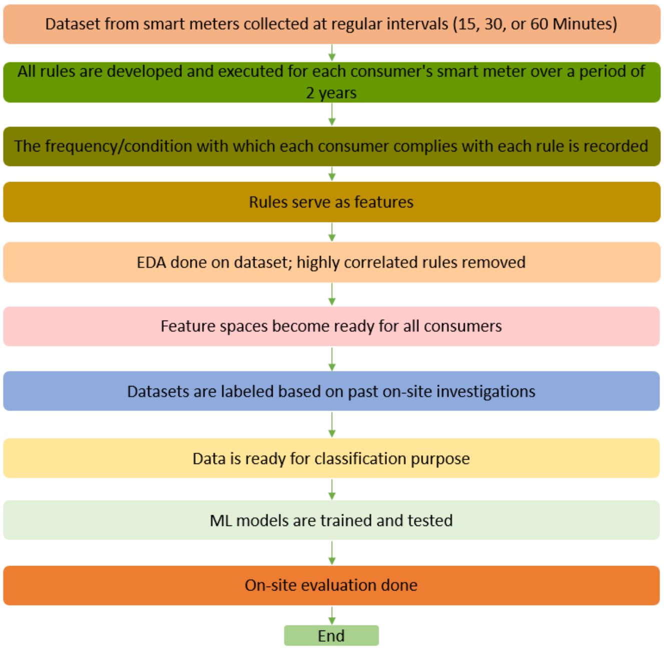



**Figure 1.** Flowchart of research methodology (Araştırma metodolojisinin akış şeması)

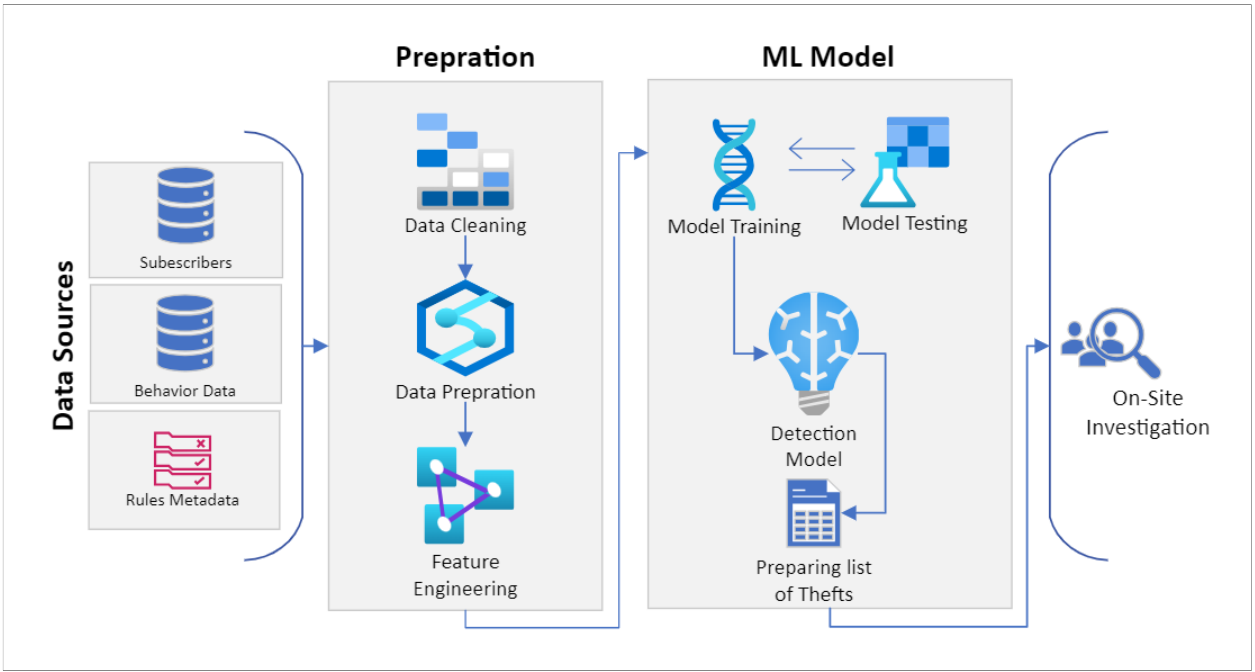



**Figure 2.** The block diagram of proposed method (Önerilen yöntemin blok diyagramı)

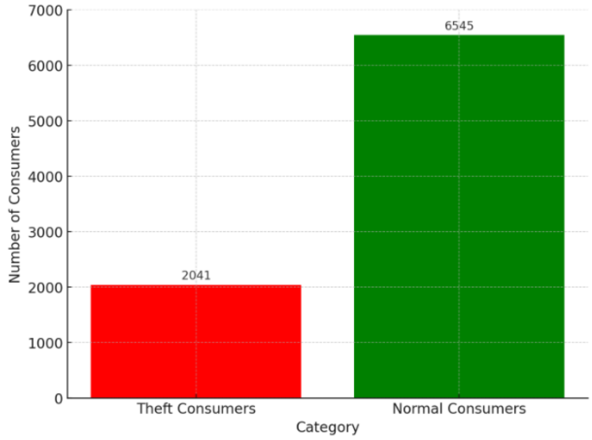
### 3.2.1 FEATURE EXPLANATION (ÖZELLİK AÇIKLAMASI)

The detailed explanation of the nine features extracted from section 3.1 is informed by expert experience, revealing their lack of intercorrelation. This insight leads to the creation of distinct feature spaces for individual consumers. Utilizing their historical data and conducting on-site inspections, consumers are then categorized as either fraudulent or normal.

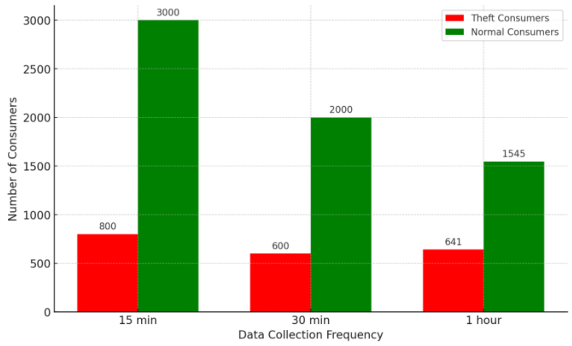
### 3.2.2 DATA PREPARATION (VERİ HAZIRLAMA)

This study uses electricity consumption data of UEDAS (Uludağ Elektrik Dağıtım A.Ş.) in Turkey. The data set consists of business consumers of electricity (8,586), differentiating them into two distinct categories: 2,041 theft consumers and 6,545 normal consumers. Figure 3 shows the number of consumers and the load profile collection frequencies. To prepare a dataset as input for machine learning models, we examined the behavior of each consumer and applied the 14 rules to each consumer individually. Subsequently, we constructed feature spaces for all consumers based on the results of these rule evaluations.

A correlation analysis was performed to look at the correlations between the 14 features once the feature spaces were developed. Cases where the correlation coefficients were higher than the cutoff of 0.5 were found by this study. Considering the importance of these high correlation values, professional advice was sought to decide on the best course of action. Experts agreed that these relationships could not be ignored because of their possible importance and influence. It was therefore suggested that all characteristics showing a correlation coefficient more than 0.5 be kept for additional examination and thought out in the research. This methodology guarantees an exhaustive assessment of all pertinent and interconnected characteristics, thus augmenting the resilience and dependability of the complete study. Subsequent to rigorous expert scrutiny, a selection of features exhibiting high intercorrelations were judiciously excised from the dataset. Consequently, a refined subset of nine features was retained, deemed most pertinent for serving as inputs to the machine learning model. Figure 4 presents a heatmap illustrating the correlations among the various features.




(a)




(b)

**Figure 3.** Distribution of electricity consumers (a) electricity consumer categories (b) electricity consumer categories and data collection frequencies (Elektrik tüketicilerinin dağılımı (a) elektrik tüketicisi kategorileri (b) elektrik tüketicisi kategorileri ve veri toplama sıklıkları)

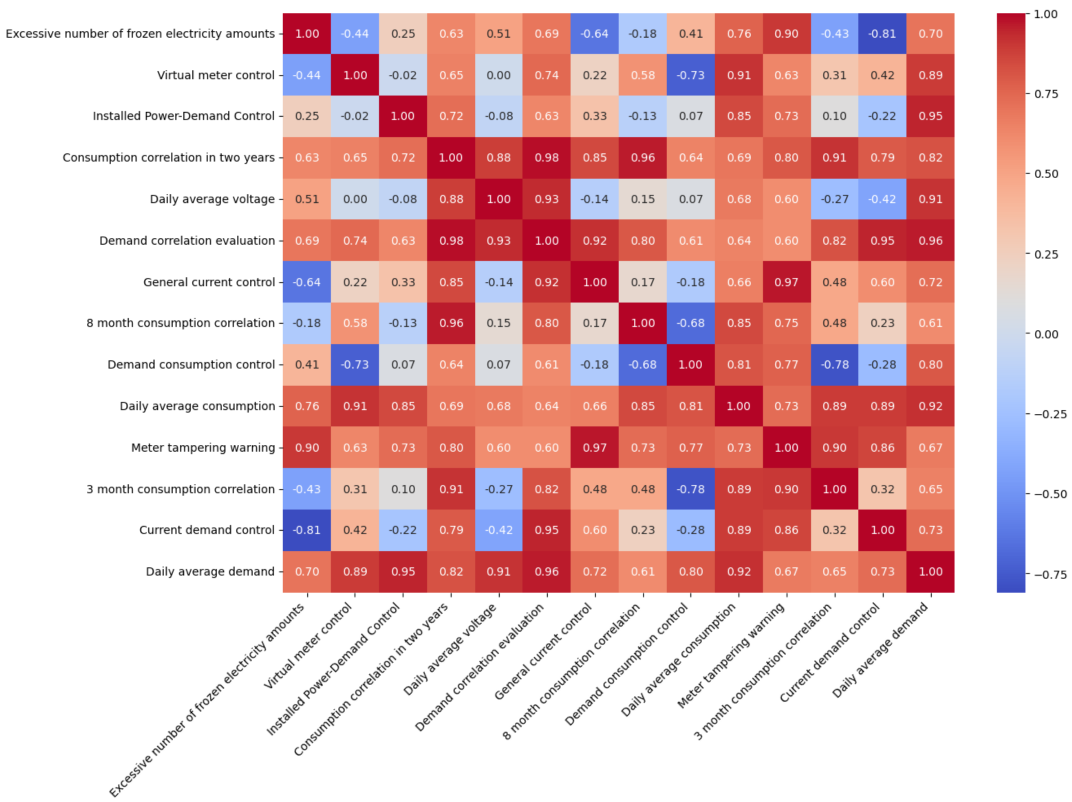



**Figure 4**. Heatmap of correlations among features (Özellikler arasındaki korelasyonların ısı haritası)

Features are categorized as either categorical or numerical. Categorical features are represented as binary values, 0 or 1, to indicate specific conditions or states. Numerical features, on the other hand, are scaled using a standard scaler, adjusting their values to fall within a range of 0 to 1. This standardization ensures consistency and comparability across different scales and measurements. Table 2 shows the extracted data after removing the unwanted features.

**Table 2**. Extracted features (Çıkarılan özellikler)

| Feature | Data Type | Description |
|---|---|---|
| 1.Meter Tampering Warning | Categorical | Assign a value of 1 if a meter tamper alarm is received and consumption has decreased, otherwise assign 0 |
| 2.Virtual Meter Control | Categorical | Assign a value of 1 if the virtual meter control is correct, otherwise assign 0 |
| 3.Daily Average Voltage | Numerical | Record and count the number of times the voltage surpasses a predetermined threshold set by experts. |
| 4.Excessive Number of Frozen Electricity Amounts | Numerical | Record the number of times freezing conditions are observed |
| 5.General Current Control | Numerical | Count and record the number of times imbalances are detected. |
| 6.Installed Power-Demand Control | Numerical | Record the number of times the installed power capacity surpasses the demand |
| 7.Demand Correlation Evaluation | Numerical | Record values that fall within a range of 0 to 1 |
| 8.Demand Consumption Control | Numerical | Calculate and record the total work hours for each sector |
| 9.Current-Demand Control | Numerical | Check for the existence of demand in any phase and record the number of times demand is lacking |

This structured approach to feature selection and preprocessing facilitates a more accurate and efficient analysis, enabling better insights and decision-making based on the data.

In other words, we analyzed the behavior of each consumer using the nine rules and used the outcomes of these rule-based evaluations to create feature sets or feature spaces for every consumer. These features can then be used as input data for machine learning models. Experts assisted in labeling the data to transform it into a classification problem. Table 3 presents the descriptive statistical metrics for the aforementioned nine features, offering a comprehensive overview of their distributional characteristics.

**Table 3.** Descriptive statistical values for the gained data (Elde edilen veriler için tanımlayıcı istatistiksel değerler)

| Parameters | Mean | Median | Mode | Minimum | Maximum | Std. Deviation | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.001 | 0 | 0 | 0 | 2 | 0.03 | 33.19 | 1227.98 |
| 2 | 0.07 | 0 | 0 | 0 | 3 | 0.26 | 3.35 | 12.67 |
| 3 | 230.86 | 231 | 230 | 0 | 447 | 10.95 | -12.99 | 266.04 |
| 4 | 0.06 | 0 | 0 | 0 | 9 | 0.39 | 9.67 | 125.80 |
| 5 | 0.40 | 0 | 0 | 0 | 29 | 2.4 | 6.98 | 55.25 |
| 6 | 0.48 | 0 | 0 | 0 | 10 | 0.3 | 10.02 | 152.06 |
| 7 | 8.1 | 4.1 | 0 | 0 | 804.1 | 11.24 | 2.91 | 17.45 |
| 8 | 0.059 | 0 | 0 | 0 | 3 | 0.23 | 3.91 | 17.51 |
| 9 | 0.005 | 0 | 0 | 0 | 11 | 0.19 | 35.89 | 1402.02 |

### 3.2.3 EVALUATION METRICS
(DEĞERLENDIRME ÖLÇÜTLERİ)

Electricity theft detection presents a challenge within the domain of imbalanced datasets. In this scenario, the dataset heavily leans towards one specific outcome of the target variable, leaving the other outcome(s) underrepresented. It's worth noting that our primary focus is on the less common outcome. Consequently, the choice of an appropriate evaluation metric becomes crucial.

The majority of consumers are not involved in theft (True Negatives), while a small number are potential thieves (True Positives). Opting for accuracy as the evaluation metric would be ill-advised since the results would be skewed towards the more prevalent class, which is True Negatives. Instead, what we need is a metric that provides a comprehensive understanding of both the actual number of thieves (recall) and the actual number of predicted thieves (precision), considering both aspects together. To address this need, we utilize the F-measure, which combines precision and recall into a single metric. In our research, we employ recall, precision, and the F-measure as our performance evaluation metrics. Equations 5-7 within our work precisely define these terms.

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$F1 - score = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \quad (7)$$

## 4. RESULTS AND DISCUSSIONS (BULGULAR VE TARTIŞMALAR)

In this segment, we establish the credibility of our research by conducting rigorous evaluations within the Python 3.10 environment. Our experimentation took place on a Windows Server running a 64-bit operating system, powered by robust hardware featuring an InteI) processor clocking in at 2.8 GHz and an ample 32 GB of RAM. The foundation of our analysis relies on the utilization of nine carefully selected features that serve as the basis for assessing the performance of various machine learning classifiers, yielding valuable insights into their effectiveness. Our evaluation process encompasses the computation of essential performance metrics, including precision, recall, and F-measure. These metrics provide a comprehensive perspective on the classifiers capabilities and their ability to distinguish between different classes. For clarity and reference, we present metrices summarizing the outcomes of all classifiers in Table 4 and Figure 5. This matrix encapsulates vital information regarding the models prediction accuracy and misclassification tendencies, contributing to a comprehensive assessment of their overall performance.

**Table 4.** Recall, F1-Score and Precision of all classifiers (Tüm sınıflandırıcıların Recall, F1-Score ve Precision değerleri)

| Classifier Type | Classifiers | Recall | F1-Score | Precision |
|---|---|---|---|---|
| Ensemble Methods | RandomForest | 0.93 | 0.65 | 0.50 |
| | AdaBoost | 0.87 | 0.65 | 0.52 |
| | LGBoost | 0.85 | 0.67 | 0.55 |
| | XGBoost | 0.74 | 0.56 | 0.45 |
| Support Vector Machine | LinearSVC | 0.89 | 0.59 | 0.44 |
| Decision Tree | DecisionTreeClassifier | 0.90 | 0.61 | 0.46 |
| Nearest Neighbors | KNeigboursClassifier | 0.81 | 0.67 | 0.56 |
| Neural Network | MLPClassifier | 0.84 | 0.62 | 0.50 |

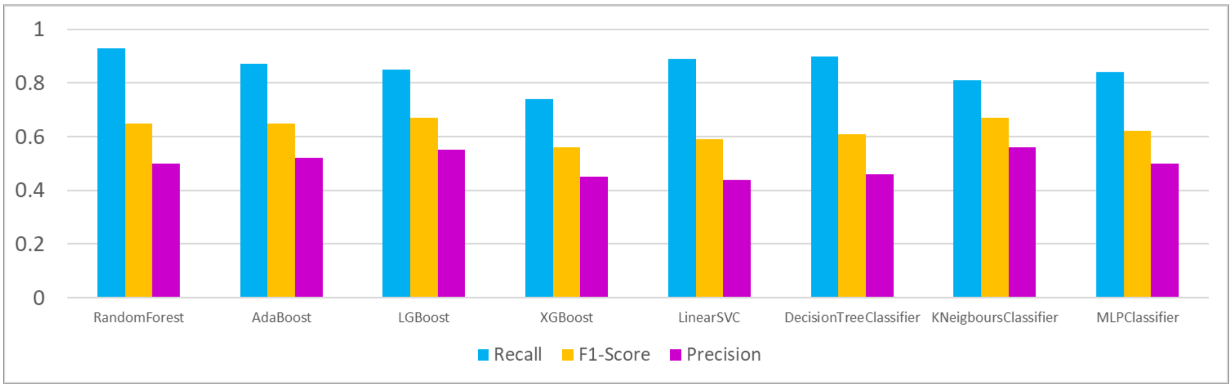



**Figure 5.** Recall, F1-Score and Precision of all classifiers (Tüm sınıflandırıcıların Recall, F1-Score ve Precision değerleri)

In the quest for the optimal classifier, we have conducted a comprehensive analysis of various machine learning models, each with its own unique strengths and weaknesses. Table 4 summarizes the key performance metrics, including recall, F1-score, and precision, for each classifier type:

Ensemble Methods have shown promising results, with Random Forest leading the pack in terms of recall (0.93). However, it is worth noting that AdaBoost, LGBoost, and XGBoost also exhibit respectable performance metrics, making them valuable contenders in our classification task. Support Vector Machine (LinearSVC) demonstrates a commendable recall of 0.89, although it falls short in terms of F1-score and precision, indicating potential room for improvement. Decision Tree classifiers, represented by Decision Tree Classifier, provide a reliable balance between recall (0.90) and F1-score (0.61), making them a competitive choice in our analysis. When it comes to Nearest Neighbor using K-Neigbours Classifier, we see a good recall (0.81) and a high F1-score (0.67), suggesting a strong ability to correctly identify positive cases in our dataset. Finally, the Neural Network model, implemented as MLP Classifier, presents a recall of 0.84, with a reasonable F1-score and precision, making it a versatile choice for our classification problem.

Ultimately, the choice of classifier should depend on the specific requirements and constraints of the task at hand. Ensemble methods like Random Forest and AdaBoost might be preferred for maximizing recall, while Decision Tree and Nearest Neighbors offer a balanced trade-off between recall and F1-score. The Support Vector Machine and Neural Network classifiers also provide competitive options for different use cases. To make a final decision, further experimentation and consideration of the specific goals and data characteristics are necessary.

In addition to our comprehensive analysis of machine learning classifiers, our research introduces three novel aspects:

- Expert-Driven Rule-Based Features: We enrich our models with nine expert-defined rules, adding depth and precision beyond typical consumption data-based approaches. This strategy enhances interpretability and aligns with current trends in machine learning.
- On-Site Investigation Integration: Our methodology includes real-world data from on-site investigations, ensuring practical relevance and applicability in real-life scenarios, a feature often missing in theoretical models.
- Real vs. Synthetic Theft Patterns: Unlike common practices using synthetic data, our study focuses on genuine theft patterns, offering a more accurate representation of real-world electricity theft scenarios, significantly improving the reliability and practicality of our detection models.

These innovative elements provide our research with a unique edge in the field of electricity theft detection, combining theoretical robustness with practical applicability.

In electricity theft studies, various methodologies and models have been employed to enhance detection accuracy while preserving privacy. Richardson et al. [50] utilized consumption data with an SVM model, injecting anomalies in feeders to detect irregularities. This approach was particularly focused on privacy preservation. In contrast, Figueroa et al. [9] enhanced the SVM model for datasets that are not balanced, addressing the challenge of skewed data distributions. Qu et al. [51] adopted the Random Forest method, with a specific focus on handling class imbalance issues. They employed the k-means and SMOTE methods for oversampling before applying the classifier, demonstrating an innovative approach to improving data representation. Nagi et al. [52] utilized expert-written rules to filter output results from the SVM model. This approach leveraged domain expertise for more precise anomaly detection. Punmiya and Choe [53] explored the use of a Gradient Boosting classifier, where synthetic theft cases were generated based on historical theft data. This method was compared against various algorithms like SVM, Backpropagation Neural Network (BPNN), Extreme Learning Machine (ELM), Deep ELM (DELM), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), AdaBoost, Naïve Bayes (NB), and k-Nearest Neighbors (kNNs). The comparison aimed to evaluate which algorithms could detect electricity theft with higher accuracy or lower False Positive Rate (FPR).

It is noteworthy that in the above-mentioned studies, there appears to be a lack of on-site investigations and a sufficient number of real theft datasets. This limitation suggests an opportunity for future research to incorporate more comprehensive data and validation methods in electricity theft detection studies.

## 5. CONCLUSIONS (SONUÇLAR)

In conclusion, the realm of electricity theft detection is inundated with various supervised and unsupervised models, all claiming efficacy in identifying theft based on consumption patterns. However, real-world complexities such as sector variations, consumer dynamics (such as hiring new staff, acquiring new devices, vacancies, and changing business structures without altering power contracts), economic downturns, and other factors render relying solely on consumption data impractical for on-site inspections.

Our study addresses this challenge by emphasizing the significance of data preprocessing and post-processing, conducted with expert guidance, to mitigate the impact of false positive detections. Unlike many existing approaches, our methodology extended beyond consumption analysis. We meticulously crafted features to encapsulate theft indicators, employing them as inputs for our models. Subsequently, on-site inspections were carried out, enhancing the accuracy of our findings. Among the models tested, Random Forest emerged as the standout performer, particularly due to its robustness in minimizing false positives and its strength in reducing overfitting—a key advantage that helped us to detect more instances of electricity theft accurately during on-site investigations. Its effectiveness is further enhanced by its ability to adeptly handle mixed numeric and categorical features without the need for extensive preprocessing or scaling, making it exceptionally versatile across different data types. Additionally, Random Forest exhibits remarkable resilience when dealing with imbalanced data, ensuring that minority classes are adequately represented. This model also benefits from requiring less hyperparameter tuning compared to other algorithms, making it both a powerful and user-friendly option for tackling complex classification tasks. By leveraging Random Forest, we achieved a substantial success rate, detecting 77% of anomalies as instances of electricity theft. This accomplishment underscores the importance of not only the choice of the model but also the thoughtful integration of domain knowledge and comprehensive feature engineering, which collectively enhance the effectiveness of theft detection efforts.

In the future work, we aim to expand our approach by incorporating additional features derived from elaborated smart meter datasets. This will involve analyzing more granular data points to better understand consumption patterns. The goal is to further decrease the rate of false positives, which will be instrumental in reducing the costs associated with on-site investigations. By refining our detection algorithms and integrating more detailed data, we hope to increase the efficiency and accuracy of electricity theft detection, providing significant cost savings for companies and improving the overall effectiveness of theft identification.

## ACKNOWLEDGEMENTS (TEŞEKKÜR)

This study was conducted as a part of the project titled "Technical and Non-Technical Losses Estimation and Warning System for Transformers and Subscribers" with the project number 7210765 and supported by TÜBİTAK.

**DECLARATION OF ETHICAL STANDARDS** (ETİK STANDARTLARIN BEYANI)

The authors of this article declares that the materials and methods they use in their work do not require ethical committee approval and/or legal-specific permission.

Bu makalenin yazarları çalışmalarında kullandıkları materyal ve yöntemlerin etik kurul izni ve/veya yasal-özel bir izin gerektirmediğini beyan ederler.

**AUTHORS' CONTRIBUTIONS** (YAZARLARIN KATKILARI)

***Sheyda BAHRAMİ:*** She conducted the literature review, software development and writing process.

Literatür taraması, yazılım geliştirme ve makale yazım sürecini yürütmüştür.

***Erol YUMUK:*** He conducted the research, data collection and control processes.

Araştırma, veri toplama ve kontrol süreçlerini yürütmüştür.

***Alper KEREM:*** He conducted the literature review, research, editing and consultancy.

Literatür taraması, araştırma, düzenleme ve danışmanlık görevlerini yürütmüştür.

***Beytullah TOPCU:*** He conducted the research, software development and data collection.

Araştırma, yazılım geliştirme ve veri toplama çalışmalarını yürütmüştür.

***Ahmetcan KAYA:*** He conducted the research, software development and data collection.

Araştırma, yazılım geliştirme ve veri toplama çalışmalarını yürütmüştür.

**CONFLICT OF INTEREST** (ÇIKAR ÇATIŞMASI)

There is no conflict of interest in this study.

Bu çalışmada herhangi bir çıkar çatışması yoktur.

**REFERENCES** (KAYNAKLAR)

[1] Emadaleslami, M., Haghifam, M. R., & Zangiabadi, M. A two-stage approach to electricity theft detection in AMI using deep learning. International Journal of Electrical Power & Energy Systems, 150, 109088, (2023).

[2] Markovska, M., Gerazov, B., Zlatkova, A., & Taskovski, D. (2023, June). Electricity Theft Detection Based on Temporal Convolutional Networks with Self-Attention. In 202[3] 30th International Conference on Systems, Signals and Image Processing (IWSSIP) (pp. 1-5). IEEE.

[3] Taha, K. Semi-supervised and un-supervised clustering: A review and experimental evaluation. Information Systems, 102178, (2023).

[4] Zheng Z, Yang Y, Niu X, Dai HN, Zhou Y. Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids. IEEE Transactions on Industrial Informatics. 2017 Dec 21;14(4):1606-15.

[5] Himeur Y, Ghanem K, Alsalemi A, Bensaali F, Amira A. Artificial intelligence-based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives. Applied Energy. Apr 1;287:116601, (2021).

[6] Oprea SV, Bâra A. Feature engineering solution with structured query language analytic functions in detecting electricity frauds using machine learning. Scientific Reports. 28; 12(1):3257, (2022).

[7] Monedero I, Biscarri F, León C, Guerrero JI, Biscarri J, Millán R. Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees. International Journal of Electrical Power & Energy Systems. 34(1):90-8, (2012).

[8] Long H, Chen C, Gu W, Xie J, Wang Z, Li G. A data-driven combined algorithm for abnormal power loss detection in the distribution network. IEEE Access. 8:24675-86, (2020).

[9] Figueroa G, Chen YS, Avila N, Chu CC. Improved practices in machine learning algorithms for NTL detection with imbalanced data. In2017 IEEE Power & Energy Society General Meeting 2017 Jul 16 (pp. 1-5). IEEE.

[10] Krishna VB, Gunter CA, Sanders WH. Evaluating detectors on optimal attack vectors that enable electricity theft and DER fraud. IEEE Journal of Selected Topics in Signal Processing. 12(4):790-805, (2018).

[11] Razavi R, Gharipour A, Fleury M, Akpan IJ. A practical feature-engineering framework for electricity theft detection in smart grids. Applied energy. 238:481-94, (2019).

[12] Peng Y, Yang Y, Xu Y, Xue Y, Song R, Kang J, Zhao H. Electricity theft detection in AMI based on clustering and local outlier factor. IEEE Access. 2021 Jul 28;9:107250-9.

[13] Jindal A, Schaeffer-Filho A, Marnerides AK, Smith P, Mauthe A, Granville L. Tackling energy theft in smart grids through data-driven analysis. In2020 International Conference on Computing, Networking and Communications (ICNC) 2020 Feb 17 (pp. 410-414). IEEE.

[14] Angelos EW, Saavedra OR, Cortés OA, De Souza AN. Detection and identification of abnormalities in customer consumptions in power distribution systems. IEEE Transactions on Power Delivery. 26(4):2436-42, (2011).

[15] Blazakis KV, Kapetanakis TN, Stavrakakis GS. Effective electricity theft detection in power distribution grids using an adaptive neuro fuzzy inference system. Energies. 13(12):3110, (2020).

[16] Fernandes SE, Pereira DR, Ramos CC, Souza AN, Gastaldello DS, Papa JP. A probabilistic optimum-path forest classifier for non-technical losses detection. IEEE Transactions on Smart Grid. 10(3):3226-35, (2018).

[17] Ramos CC, Souza AN, Papa JP, Falcao AX. Fast non-technical losses identification through optimum-path forest. In2009 15th International Conference on Intelligent System Applications to Power Systems 2009 Nov 8 (pp. 1-5). IEEE.

[18] Ramos CC, Souza AN, Chiachia G, Falcão AX, Papa JP. A novel algorithm for feature selection using harmony search and its application for non-technical losses detection. Computers & Electrical Engineering. 37(6):886-94, (2011).

[19] Ramos CC, de Sousa AN, Papa JP, Falcao AX. A new approach for nontechnical losses detection based on optimum-path forest. IEEE Transactions on Power Systems. 26(1):181-9, (2010).

[20] Ramos CC, de Souza AN, Falcao AX, Papa JP. New insights on nontechnical losses characterization through evolutionary-based feature selection. IEEE Transactions on Power Delivery. 27(1):140-6, (2011).

[21] Xiao F, Ai Q. Electricity theft detection in smart grid using random matrix theory. IET Generation, Transmission & Distribution. 12(2):371-8, (2018).

[22] Depuru SS, Wang L, Devabhaktuni V, Green RC. High performance computing for detection of electricity theft. International Journal of Electrical Power & Energy Systems. 47:21-30, (2013).

[23] Jain S, Choksi KA, Pindoriya NM. Rule-based classification of energy theft and anomalies in consumers load demand profile. IET Smart Grid. 2(4):612-24, (2019).

[24] Biswas PP, Cai H, Zhou B, Chen B, Mashima D, Zheng VW. Electricity theft pinpointing through correlation analysis of master and individual meter readings. IEEE Transactions on Smart Grid. 11(4):3031-42, (2019).

[25] Shah AL, Mesbah W, Al-Awami AT. An algorithm for accurate detection and correction of technical and nontechnical losses using smart metering. IEEE Transactions on Instrumentation and Measurement. 69(11):8809-20, (2020).

[26] Zheng K, Chen Q, Wang Y, Kang C, Xia Q. A novel combined data-driven approach for electricity theft detection. IEEE Transactions on Industrial Informatics. 15(3):1809-19, (2018).

[27] Tao J, Michailidis G. A statistical framework for detecting electricity theft activities in smart grid distribution networks. IEEE Journal on Selected Areas in Communications. 38(1):205-16, (2019).

[28] Xiang M, Rao H, Tan T, Wang Z, Ma Y. Abnormal behaviour analysis algorithm for electricity consumption based on density clustering. The Journal of Engineering. 2019(10):7250-5, (2019).

[29] Kharal AY, Khalid HA, Gastli A, Guerrero JM. A novel features-based multivariate Gaussian distribution method for the fraudulent consumers detection in the power utilities of developing countries. IEEE Access. 9:81057-67, (2021).

[30] Hock D, Kappes M, Ghita B. Using multiple data sources to detect manipulated electricity meter by an entropy-inspired metric. Sustainable Energy, Grids and Networks. 21:100290, (2020).

[31] Singh SK, Bose R, Joshi A. Entropy-based electricity theft detection in AMI network. IET Cyber-Physical Systems: Theory & Applications. 3(2):99-105, (2018).

[32] Jaiswal S, Ballal MS. Fuzzy inference based electricity theft prevention system to restrict direct tapping over distribution line. Journal of Electrical Engineering & Technology. 15:1095-106, (2020).

[33] Aligholian A, Farajollahi M, Mohsenian-Rad H. Unsupervised learning for online abnormality detection in smart meter data. In2019 IEEE Power

& Energy Society General Meeting (PESGM) 2019 Aug 4 (pp. 1-5). IEEE.

[34] Feng Z, Huang J, Tang WH, Shahidehpour M. Data mining for abnormal power consumption pattern detection based on local matrix reconstruction. International Journal of Electrical Power & Energy Systems. 123:106315, (2020).

[35] Lu X, Zhou Y, Wang Z, Yi Y, Feng L, Wang F. Knowledge embedded semi-supervised deep learning for detecting non-technical losses in the smart grid. Energies. 12(18):3452, (2019).

[36] Hu T, Guo Q, Shen X, Sun H, Wu R, Xi H. Utilizing unlabeled data to detect electricity fraud in AMI: A semisupervised deep learning approach. IEEE transactions on neural networks and learning systems. 30(11):3287-99, (2019).

[37] Lu X, Zhou Y, Wang Z, Yi Y, Feng L, Wang F. Knowledge embedded semi-supervised deep learning for detecting non-technical losses in the smart grid. Energies. 12(18):3452, (2019).

[38] Javaid N, Gul H, Baig S, Shehzad F, Xia C, Guan L, Sultana T. Using GANCNN and ERNET for detection of non technical losses to secure smart grids. IEEE Access. 9:98679-700, (2021).

[39] Ismail M, Shaaban MF, Naidu M, Serpedin E. Deep learning detection of electricity theft cyber-attacks in renewable distributed generation. IEEE Transactions on Smart Grid. 11(4):3428-37, (2020).

[40] Charwand M, Gitizadeh M, Siano P, Chicco G, Moshavash Z. Clustering of electrical load patterns and time periods using uncertainty-based multi-level amplitude thresholding. International Journal of Electrical Power & Energy Systems. 117:105624, (2020).

[41] Xia R, Gao Y, Zhu Y, Gu D, Wang J. An attention-based wide and deep CNN with dilated convolutions for detecting electricity theft considering imbalanced data. Electric Power Systems Research. 214:108886, (2023).

[42] Messinis GM, Hatziargyriou ND. Unsupervised classification for non-technical loss detection. In2018 Power Systems Computation Conference (PSCC) 2018 Jun 11 (pp. 1-7). IEEE.

[43] Breiman L. (2001) Random forests. Machine learning. Oct; 45:5-32.

[44] Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences. 55(1):119-39, (1997).

[45] Ghori KM, Abbasi RA, Awais M, Imran M, Ullah A, Szathmary L. Performance analysis of different types of machine learning classifiers for non-technical loss detection. IEEE Access. 26;8:16033-48,(2019).

[46] Raza M, Awais M, Ellahi W, Aslam N, Nguyen HX, Le-Minh H. Diagnosis and monitoring of Alzheimer's patients using classical and deep learning techniques. Expert Systems with Applications. 136:353-64, (2019).

[47] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 323(6088):533-6, (1986).

[48] Mujeeb S, Javaid N, Ahmed A, Gulfam SM, Qasim U, Shafiq M, Choi JG. Electricity theft detection with automatic labeling and enhanced RUSBoost classification using differential evolution and Jaya algorithm. IEEE Access. 9:128521-39, (2021).

[49] Khan ZA, Adil M, Javaid N, Saqib MN, Shafiq M, Choi JG. Electricity theft detection using supervised learning techniques on smart meter data. Sustainability. 12(19):8023, (2020).

[50] Richardson C, Race N, Smith P. A privacy preserving approach to energy theft detection in smart grids. In2016 IEEE International Smart Cities Conference (ISC2) 2016 Sep 12 (pp. 1-4). IEEE.

[51] Qu Z, Li H, Wang Y, Zhang J, Abu-Siada A, Yao Y. Detection of electricity theft behavior based on improved synthetic minority oversampling technique and random forest classifier. Energies. 13(8):2039, (2020).

[52] Nagi J, Yap KS, Tiong SK, Ahmed SK, Nagi F. Improving SVM-based nontechnical loss detection in power utility using the fuzzy inference system. IEEE Transactions on power delivery. 26(2):1284-5, (2011).

[53] Punmiya R, Choe S. Energy theft detection using gradient boosting theft detector with feature engineering based preprocessing. IEEE Transactions on Smart Grid. 10(2):2326-9, (2019).