IMPROVING FRAUD DETECTION IN NATURAL GAS DISTRIBUTION BY COMBINING TRADITIONAL METHODS WITH SPECIFIC DATA MINING ALGORITHMS

Silviu GRESOI[1], Ștefan MOCANU[2], Ioana FĂGĂRĂȘAN[3], Grigore STAMATESCU[4]

Energy systems, either based on electricity or natural gas, are omnipresent regardless of the financial status of a country or an area. Due to a variety of causes (such as depletion of resources, natural calamities or speculative behaviors of producers and distributors – not to mention geo-political problems) prices for electricity or natural gas may significantly vary even on short terms. Unfortunately, these often lead to situations where theft becomes a reality, and some end users find solutions to defraud energy systems. This paper is based on an actual study which aimed to identify geographic areas that exhibit high risks of energy fraud, in particular natural gas thefts. In fact, as it will be presented in the paper, results are very encouraging, a very significant number of frauds were identified by using relatively cheap methods based on data analysis and data mining. In practice, this leads to minimizing losses for distributing company (both in natural gas and human efforts) and maintaining prices for the end user at levels affected only by market behavior.

Keywords: data analysis; data cleaning; data aggregation; data mining; decision-making; fraud detection; fraud marker; anomaly detection; scoring; related analysis

1. Introduction

According to Oxford Dictionary, fraud is defined as “a type of dishonesty calculated to obtain advantage, generally financial advantage, by some wrongful means (a tort or crime)”. Cambridge Dictionary provides a similar definition “fraud [is] the crime of obtaining money or property by deceiving people”.

[1] PhD student, Faculty of Automatic Control and Computer Science, University POLITEHNICA of Bucharest, Romania, e-mail: silviu@4help.ro
[2] Professor, Faculty of Automatic Control and Computer Science, University POLITEHNICA of Bucharest, Romania, e-mail: stefan.mocanu@upb.ro
[3] Professor, Faculty of Automatic Control and Computer Science, University POLITEHNICA of Bucharest, Romania, e-mail: ioana.fagarasan@upb.ro
[4] Professor, Faculty of Automatic Control and Computer Science, University POLITEHNICA of Bucharest, Romania, e-mail: grigore.stamatescu@upb.ro

Both dictionaries, and many similar others, emphasize the deceiving part of a fraud but also relate the fraud primarily to the action obtaining money directly and only secondarily to the indirect actions that can generate money.

It is worth observing that, on Internet, the vast majority of resources are related to financial fraud detection. In paper [1] authors present an interesting review dedicated to how Machine Learning can be used for financial fraud detection. Other authors [2] extend their investigation over fraud targets to telecommunication systems, health insurance, car insurance and other areas. Although, in the past, statistical methods were used for fraud detection [3], nowadays detection methods are almost exclusively based on Data Mining and Machine Learning.

Energetic systems and the energy industry are frequent targets of frauds. Although natural gas transport and distribution networks are omnipresent, most researchers approach only fraud detection in electric systems regardless of the method. In paper [4] authors present an extended study of fraud detection in electric energy systems and mention that results can be extended to water and natural gas transport systems, but they provide no detail about how this extension can be done taking into consideration the particularities of each system. There are few specific studies on fraud detection in natural gas systems with practical implementations available.

The rest of the paper is organized as follows: Section 2 presents current approaches in fraud detection using Data Minning and Machine Learning; Section 3 is dedicated to the objectives, methodology and architecture of proposed system; Section 4 presents actual results and some discussions while in Section 5 some conclusions are drawn.

2. Current approaches in fraud detection using data mining and machine learning

As it was briefly shown in the previous section, fraud is present in, but not limited to, all areas that imply commercial or financial exchanges. Some forms of non-financial fraud, such as intellectual property fraud, plagiarism, or identity fraud can be easily identified and proven since the purpose of the fraud becomes public.

Fraud detection is strongly related to the identification of anomalies. We can define anomaly as a behavior or event that is not normal or does not fall into a pattern. The “normal” (or the pattern) is built by analyzing a considerable amount of data from the past and defining the expected behavior using similarity, mean, median, standard deviation, regression and so on. In paper [5] the authors dedicated a huge amount of work to analyzing 290 research papers and provide a highly relevant review over the use of Machine Learning in Anomaly Detection.

29 Machine Learning algorithms/models were identified and depicted and recommendations for interested researchers were provided.

The direct connection between anomaly and fraud is explored and depicted in many recent papers [6][7][8].

Papers [9] and [10] both investigate anomaly detection in electricity usage in residential buildings. While anomalies can indicate a potential fraud, above mentioned studies are focused on early fault detection or suboptimal behavior of internal electric network, with direct implications over building safety or management. Paper [4], already mentioned in Introduction, presents a very interesting holistic approach that provided the fraud detection module algorithm of a patent pending system. To quote the authors, "Exponentially less raw data is required more features are generated". Various combinations between mathematical apparatus and Data Mining algorithms are performed and obtained results validate authors' work. It is worth mentioning that, in the end, authors present some weaknesses (higher rates of false positives in real field conditions, many false positives due to data mismatch, limitations due to multitude of anomalies present in electric energy systems), but all of them are generated by particularities of electric energy systems.

In fact, the electric energy system particularities allowed authors of [11] to differentiate the users into two classes, which are high and low voltage energy consumers, and use different techniques for detecting fraud. For the low voltage energy consumers, which is the typical case of private person, a hybrid data mining technique was used. Data analysis results were good, but the authors show that electrical energy companies must refine the results through direct inspection, this conclusion being similar with the results obtained in case of natural gas systems. The authors of [12] go even further and use data from field inspections to extend the fraud and non-fraud event database which, in time, leads to improving the precision of their model and also to adapt easier to new fraud techniques.

The studies analyzed so far indicate that, on the market, there aren't many (or at all) solutions dedicated to fraud detection in natural gas energy systems and this is what this study aims to cover.

3. Objectives, methodology and architecture of proposed system

In the actual case of Romania, the first step in approaching and reducing the fraudulent behavior was to determine a fraud marker calculated as a ratio between the number of detected frauds and the number of consumers. In addition, the consumers are clustered into 3 categories (named UAT), as follows: UAT I – cities with more than 20.000 consumers; UAT II – small cities with 4.000 to 20.000 consumers and UAT III – villages or small cities with 500 to 4.000 consumers. Fig. 1 presents the fraud marker in several counties in Romania.

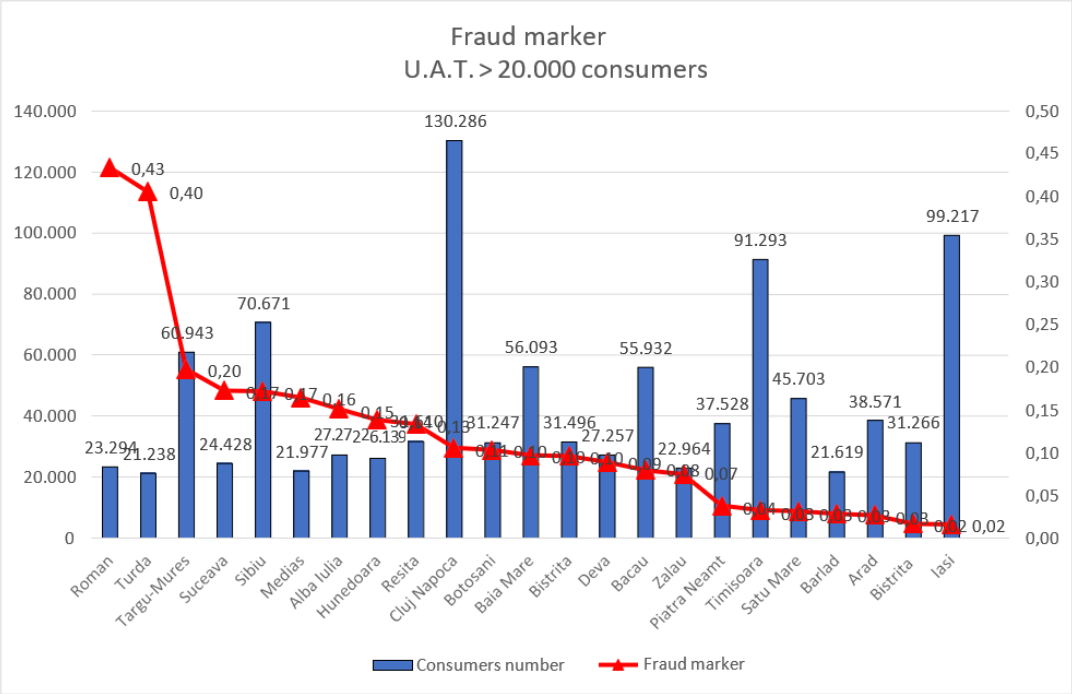

Fig. 1. Fraud marker in several UAT I (Romania)

It is easy to observe in Fig. 1 that fraud marker is very low in Iasi county (0,02). This can indicate a very low criminal behavior but, also, can indicate a very low detection rate. The management decision was to consider Iasi county as a geographic area with a high fraud potential, so field investigations were started. In parallel, data analysis was conducted.

For the beginning, a village in Iasi county was selected for the field inspection campaign. As a result of the inspections, from a total of 3226 consuming points 2000 were randomly verified and 4 frauds were detected, leading to a fraud marker of 0,2% which was considerably higher than the one presented in Fig. 1.

Starting from the success of the field inspection campaign, the next step was to continue the investigation, this time using Data Mining, profiling, and scoring methods. The 4 fraudulent customers mentioned before did not offer a solid base for the scoring. For this reason, the dataset had to be enlarged. We used 1633 non-fraud cases confirmed by the field inspection mentioned above to which we added another 1633 cases of confirmed frauds within the geographic area of the entire Iasi county. We obtained a balanced dataset of 3266 records on a period of time covering 2016-2021.

The fraud cases were also used to build patterns for fraudulent customers, some characteristics of the constructed pattern being presented in Table 1 along with a fraud type, fake metrological mark. Leu (or RON) is the official currency in Romania, and it was the equivalent of 0,2035 EUR in November 2022.

Table 1.

Characteristics of fraudulent consumer

Id	Meter Type	Fraud Type	Meter Age	Consumption Value	Private Heating Plant	Location	Duo Contract	Verified
Sql142	Natural gas	Fake metrological mark	5 years	<150 Lei ~ /Month	Yes	Rural	Yes	No

The main purpose of the current study was to correctly identify a geographic area with a high fraud risk generated by unauthorized interventions over the natural gas distribution network. The data used for analysis was provided by DELGAZ GRID [13] and E.ON Romania [14]. By using Data Mining algorithms, we aim to process data so, in the end, results will indicate with higher precision the possible fraudulent consumers within a smaller zone. This way, field inspections will be targeted, the costs with inspections will be lower and the detection efficiency will be better. The detection efficiency is defined as follows: detection efficiency = number of detected frauds / number of inspections. To increase the detection efficiency of unauthorized consumption, the inverted pyramid method will be used. In fact, the idea of extending investigations from small village to county, and possibly, in the future, to a region (for example, Moldova, in case of Romania) was borrowed from the inverted pyramid method.

The study used data collected over a period of 13 months and took into consideration several indicators: financial indicators, usage profiles, date of the last inspection, fabrication year of the meter, overdue payments, periods of lack of consumption, billing interval, GIS based location, and others. Fig. 2 presents the entire process flow, from data collection to final data matrix and then to scoring and decision.

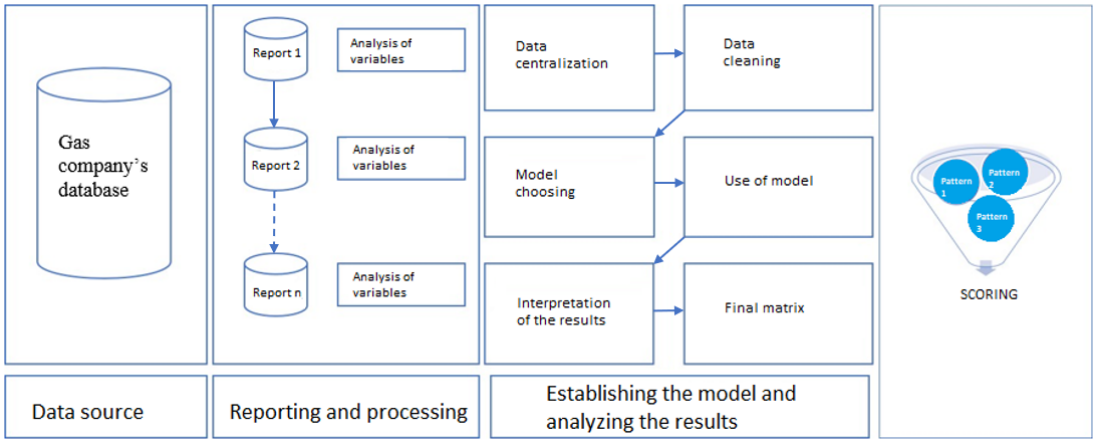

Fig. 2. Data collection and processing

Before conducting actual analysis, several steps consisting in data cleaning and data pre-processing had to be performed. All of the 3266 consumption points raw data were analyzed and unified reports were generated and grouped into:

- General information like consumer name, ID, account etc.
- Geographic information like GPS coordinates, county, street etc.
- Economic information like last payment, overdue payments, debts
- Social information like single
- Technical information like consumption type, measurement unit (m3/month)

Aside meeting some technical requirements, the collected data were subjected to an anonymization process that led to the removal of some customer identification details such as customer name, phone number, contract number etc. The general methodology was based on functional networks and specific steps were followed: (1) variables selection, (2) data filtering, (3) model choosing, (4) model analysis and (5) model evaluation. Implementation was based on WEKA [15].

Data centralization represents the first systematization of the analyzed information and consists in dividing the units into homogeneous groups, depending on the variation of one or more characteristics. An important concern resides in avoiding the introduction of erroneous information into the data matrix, which can lead to the distortion of the context of the analysis. For example, null values are eliminated; the approach is carried out through an exploratory analysis.

Data centralization involves grouping units based on multiple characteristics through recursive procedures. The first step involves selecting a primary grouping characteristic and distributing units based on its values.

Subsequent steps divide each group or subgroup further based on variations of other characteristics. For instance, installed meters can be grouped based on model, age, and recorded consumption, with each group then divided

based on intervals of the first characteristic. To avoid excessive division, it is advisable to limit the grouping features to a maximum of three or four.

Since consumption depends on a multitude of factors (area of the location, number of tenants, energy efficiency, private or company use, etc.) it is obvious that the consumption values can be extremely far apart even within a small geographical area. This led to the need to introduce an indicator called variation coefficient, where variation represents the property of a characteristic to take values in a certain interval, under the influence of time or space conditions. The variation coefficient is defined as follows: variation coefficient = (std/mean)*100, where std is standard deviation.

Table 2 presents statistical indicators of some of the variables analyzed in the study. Huge differences between variation coefficients values can be easily observed. Also, one can observe the values from column Null, which indicate the number of errors observed within the recorded data per different characteristics (variables).

Table 2.

Sample values for statistical indicators

ID	Variable	Null	Average	Minimum	Maximum	Variation Coefficient
T01	Consumption bill	122	175,47 LEI	9,34	439,43 LEI	2.125%
T02	MWh	122	1,081	0,024	213,80	1.135%
…	…	…	…	…	…	…
T60	Meter installation year	9	2018,07	2013	2021	8%

Due to above mentioned huge differences between the values of variation coefficients, data normalization was mandatory. For this study, two types of normalization were used.

i) Min-max normalization:

$$v' = \frac{v - min}{max - min} \quad \text{(Eq. 1)}$$

where:
- v’ is the new value,
- v is the old value,
- min and max are minimum and maximum value of the data.

ii) Z-score normalization:

$$v' = \frac{v - mean}{std} \quad \text{(Eq.2)}$$

where:
- v' is the new value,
- v is the old value,
- mean is the mean of available data,
- std is the standard deviation of data.

As mentioned earlier , the data set, consisting of 3266 records, was well balanced, so there were good premises for supervised learning algorithms to run in good conditions.
The main methods used in this study are similar with the ones described in [16] to which Supervised Learning was added. The primary objective of Supervised Learning was to predict frauds within a geographic area, based only on data analysis, so the number of field inspection could be as low as possible and only for the customers identified as possible fraudsters, this way increasing the detection efficiency. For training purposes, the fraud records were annotated with "1" while the non-fraud with "0". Two Supervised Learning algorithms were tested in WEKA, namely k Nearest Neighbor (kNN) and Decision Tree (DT), similar to the strategies presented in [17].

Fig. 3. Results obtained by running kNN

The dataset was shuffled and split into 2 equal subsets before running kNN (IBk in WEKA) twice. Fig. 3 shows the results for the first subset, which indicates an overlap of fraud and non-fraud cases in the low gas consumption range. The analysis will focus on consumers with consumption values below the average of

200 lei/billed monthly, resulting in 328 suspected fraudulent consumers. The second subset identified 214 possible fraudulent consumers. In total, the analysis of the complete dataset of 3266 records yielded 542 suspected fraudulent consumers, and other indicators were averaged.

Other numerical results after running the kNN are presented in Table 3. As one can observe, the obtained accuracy was a very good one, close to 99,87%.

Table 3.

Numerical results of kNN

IBk	**Minim**	**Maxim**	**Mean**	**StdDev**	**Accuracy of instances**	**Error coefficient**
Avg. consumption	0,001	1200,915	76,115	139595	99,8655%	0,1345%
Total consumption	0,011	2393,937	133,55	162,966		

Running the DT algorithm in WEKA generated the CART tree presented in Fig. 4. Same dataset was used as in kNN.

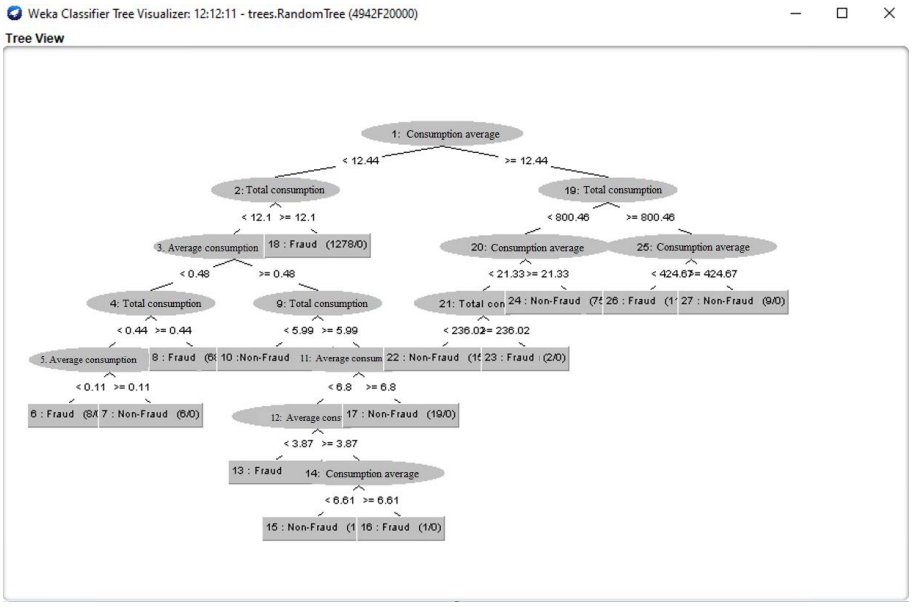

Fig. 4. The Decision Tree built using the training dataset

Numerical results obtained by running the DT are presented in Table 4. As one can observe, the accuracy obtained was not a bad one, close to 95,25%, but was lower that kNN's one. For this reason, kNN became the favorite algorithm for supervised learning in this study and based on its results field inspections for 542 customers were decided.

Table 4.

Numerical results of DT

Decision Tree	Minimum	Maximum	Mean	StdDev	Accuracy of instances	Error coefficient
Consuption average	0,001	1200,915	76,115	139,595	95,2494%	4,7506%
Consumption total	0,011	2393,937	133,55	162,966		

The use of supervised learning algorithms was made as a way to quickly identify customers suspected of fraud and to reduce field inspections. For example, at the beginning of the study, 2000 random field inspections were performed in a village with 3226 consumption points. This is the equivalent of about 62% customers being inspected. After running the data mining kNN algorithm, only 542 customers (this time out of 3266) were subject to inspections, which means about 16,6%. It is obvious that even if all of the 542 customers are non fraudulent and the inspections are made, the gas distribution company will seriously reduce costs. A rough estimate shows 3,5-4 less field inspections! The inspection manager can decide additional inspections, random or not, just to confirm the predictions made by kNN.

Analyzing consumption records predicts unauthorized interventions. Fig. 5 show data and graphs for identified customers (PF1, PF2, PF3, and PJ1) involved in unauthorized interventions in Iași County.

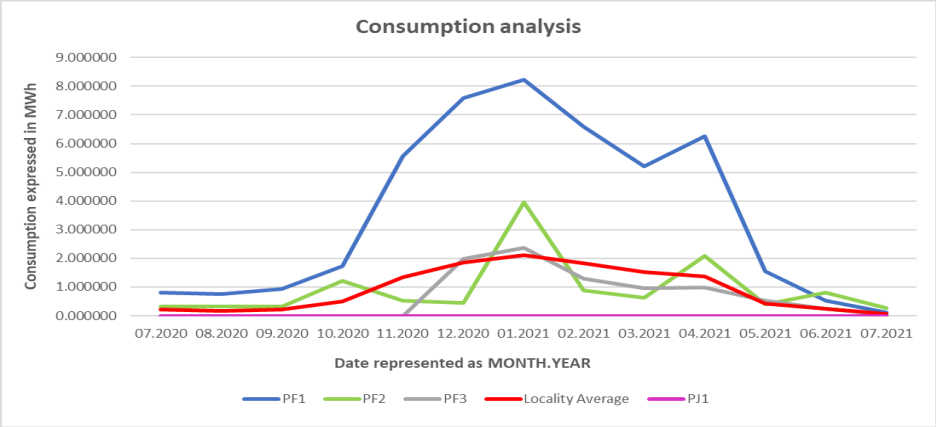

Fig. 5. Graphic representation of fraudulent consumers' consumption

Anomalies in consumption are visible for PF2, PF3, and PJ1 in Fig. 5. These customers deviate significantly from the locality's average consumption. Based on Table 4, these 3 customers had lower consumption 69.23% of the time. However, customer PF1 had unauthorized intervention but had higher actual consumption compared to the locality's average. Considering this aspect is crucial to avoid false positive results.

To perform a better analysis and to improve the detection efficiency, the final component presented in Fig. 2, scoring, was developed based on technical, economical, operational, and behavioral patterns. Scoring is the final component developed for reducing the inspections area and field inspections by refining the results of kNN. To assess fraud risk, we added a characteristic to each consumer, grouping them into 3 distinct groups based on kNN results. Table 5 shows group names, distribution criteria, customer count, and risk labels.

Table 5.

Groups and levels of risk

Group 1	Score > 15	88 consumption spots	High risk
Group 2	Score between 5 -15	211 consumption spots	Medium risk
Group 3	Score < 5	243 consumption spots	Low risk

Over 15,000 fraud records from 2010 to 2021 were analyzed to create 54 fraud profiles, each with a score index of 1 to 3 indicating low to high risk. These profiles were generated using the kMeans algorithm for clustering through Unsupervised Learning, with the value for "k" determined using the Silhouette method [18]. As customers can belong to multiple profiles, their final score is obtained by summing each profile score.

Then the final score is compared to some thresholds, as one can see in Table 5, and distributed to the appropriate group. Further details about the profiles cannot be publicly disclosed because they represent private knowledge of the companies [13][14].

The inspection manager can prioritize visits based on the level of risk, inspecting high-risk customers before moving on to medium or low-risk customers. For distribution companies with many clients, inspections can focus only on high-risk customers. Through data mining, scoring, grouping, and prioritization, inspections can be phased, starting with the 88 customers with the highest risk of unauthorized gas consumption (2.69% of the total consumers analyzed). Fig. 6 shows the geographical area subject to random inspection (a) and the locations with high, medium, and low risk that will be inspected based on the proposed methodology (b). The circled area in Fig. 6a has a radius of approximately 1.80 km, covering over 10 km2, while Fig. 6b shows easily identifiable targeted inspections.

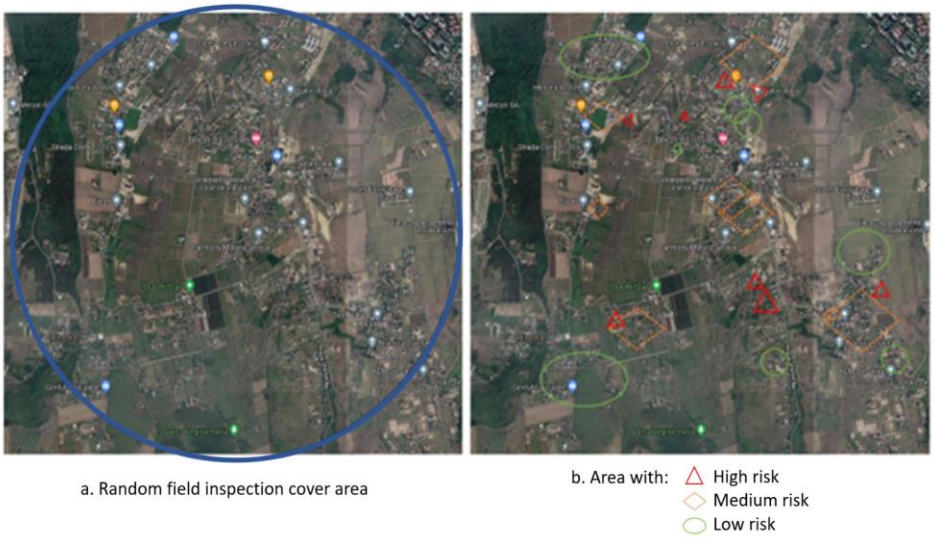

Fig. 6. Field inspection a) random b) targeted and priority based

The research study introduces another new concept that allows the expansion of the results obtained through data analysis and field inspection, a concept that will be presented, for wellfounded chronological reasons, in the next section.

5. Results and discussions

Numerical results from the previous section led to a field inspection that prioritized high, medium, and low risk customers. A confusion matrix was created to validate the data analysis, with values for True Positive, True Negative, False Positive, and False Negative recorded for fraud. Various types of fraud were detected during the inspection, ranging from meter avoidance to violent interventions. Actual results after the field inspection can be observed in Table 6 in which detection efficiency was calculated considering only the consumption spots where inspectors had access.

Table 6.

Results of field inspections based on data analysis

Category	Score	Consumption spots	Verified locations	No access	Fraud reports	Defects reports	Total reports	Efficiency
Group 1	>15	88	75	13	8	2	10	13,3%
Group 2	5 - 15	211	182	29	12	7	19	10,4%
Group 3	< 5	243	163	80	3	1	4	2,5%

It is easy to observe that detection efficiency is dramatically improved, from only 0,02% in the case of initial random field inspection. If efficiencies for all three groups are averaged, we obtain an average efficiency of 8,73% which still is much better than the initial one. Data from Table 6 can be interpreted so they cover not only actual inspections but also worse and bestcase scenarios. If we consider all locations where inspectors could not perform their check (No access points of consumption) as frauds, then we get, for Group 1 to 3, total reports of 23, 48 and 84. In this case, efficiencies will be 26,13%, 22,75% and 34,57%, with an average of 27,817%. If "No access" points of consumption are considered non-fraud, then total reports for Group 1 to 3 will not change, but the efficiencies will. In this case we will have efficiencies of 11,36%, 9% and 1,65% with an average of 7,34%. Even in this case, the results are still much better than the ones based on random inspection.

So far, the data analysis performed before the last-mentioned field inspection was confirmed by the latter. However, as mentioned at the end of Section 4, at this point a new concept was introduced in the expertise of the field of detection of unauthorized interventions at the level of practices of gas Distribution and Supply companies. This concept is represented by related analysis.

The analysis aims to identify fraudulent customers with multiple points of consumption and inspect them even if there are no indications of fraud from other sources. Although speculative, this approach has proven effective. Sensitive customer data is required for this analysis, obtained from legal sources. Table 7 presents the results of the related analysis conducted on customers identified with fraud during the field inspection.

Table 7.

Results of related analysis

Category	Verified	Fraud reports	Defects reports	Related analysis	Total	Final efficiency
Group 1	75	8	2	3	13	**17,33%**
Group 2	182	12	7	2	21	**11,54%**
Group 3	163	3	1	4	8	**4,91%**

In addition to Table 7, an additional column was introduced (and highlighted) that represents the number of new detected frauds (9). This proves that related analysis falls into inverted pyramid methods mentioned at the beginning of Section 4. By comparing detection efficiency from Table 7 with the one presented in Table 6 we can observe significant improvements in case of Group 1 and Group 3.

6. Conclusions and future developments

The methodology of narrowing the geographic area and field inspections developed in the present research study, represents a new concept in Romania among natural gas supply or distribution companies. According to [19], even at international level, the use of data analysis and data mining techniques by natural gas distribution companies is very low, which makes this study more relevant. In addition to traditional analysis methods and data mining, new concepts were introduced and validated.

This study achieved higher identification rates for unauthorized interventions, benefiting natural gas distribution companies by recovering losses. As a pilot study, it recovered 204,506 lei (~41,600 EUR) from 42 fraudulent events, averaging 1,000 EUR per detected fraud. This information aids financial departments in making projections based on customer numbers.

The research on the inverted pyramid approach for detecting unauthorized interventions in energy yielded compelling results. Contrary to conventional thinking, expanding analysis area or risk doesn't decrease detection rate. Our methodology increased detection rate from 0.02% to 8.73%. These findings impact strategies to prevent unauthorized interventions. Further research can refine the inverted pyramid approach for better effectiveness. Related analysis enhances detection of unauthorized interventions, improving accuracy. Integrating analytics and behavioral interventions benefits customers. Methodology identified rise from 33 to 42 unauthorized interventions.

The utilization of score indexes (employing pattern analysis), along with grouping and prioritization (based on the degree of risk), constitutes an individual's contribution to unauthorized intervention research. These methods have significantly enhanced the rate of identification, but their efficacy cannot be attributed to any other specialized studies.

For the future, this study will be continued with aggregated analysis of electricity-natural gas. This direction is promising, and it is expected to generate improvements in fraud detection rates. In the end, based on the results obtained and presented in this paper, we conclude that combining traditional inspection and analysis methods with Data Mining and Machine Learning is benefic for gas distribution companies in their fight to efficiently detect and stop fraudulent behavior.

R E F E R E N C E S

[1]. *Sinayobye Janvier Omar, Kiwanuka Fred, and Kaawaase Kyanda Swaib*. 2018. A state-of-the-art review of machine learning techniques for fraud detection research. In Proceedings of the 2018 International Conference on Software Engineering in Africa (SEiA '18).

Association for Computing Machinery, New York, NY, USA, 11–19. https://doi.org/10.1145/3195528.3195534

[2]. *Aisha Abdallah; Mohd Aizaini Maarof; Anazida Zainal;* Fraud detection system: A survey, Journal of Network and Computer Applications, Vol. 68, 2016, pp 90-113, https://doi.org/10.1016/j.jnca.2016.04.007

[3]. *Richard J. Bolton; David J. Hand.* "Statistical Fraud Detection: A Review." Statistical Science, 17(3) 235-255 Au-gust 2002. https://doi.org/10.1214/ss/1042727940

[4]. *Calamaro Netzah; Yuval Beck; Ran Ben Melech; Doron Shmilovitz.* "An Energy-Fraud Detection-System Capable of Distinguishing Frauds from Other Energy Flow Anomalies in an Urban Environ-ment", Sustainability 2021,13, no. 19: 10696. https://doi.org/10.3390/su131910696

[5]. *A. B. Nassif; M. A. Talib; Q. Nasir; F. M. Dakalbab*, "Machine Learning for Anomaly Detection: A Systematic Review," in IEEE Access, vol. 9, pp. 78658-78700, 2021, doi: 10.1109/ACCESS.2021.3083060.

[6]. *E. L. Paula; M. Ladeira; R. N. Carvalho; T. Marzagão,* "Deep Learning Anomaly Detection as Support Fraud Investigation in Brazilian Exports and Anti-Money Laundering", 2016 15th IEEE International Conference on Ma-chine Learning and Applications (ICMLA), Anaheim, CA, USA, 2016, pp. 954-960, doi: 10.1109/ICMLA.2016.0172.

[7]. *Yulong Pei; Fang Lyu; Werner van Ipenburg; Mykola Pechenizkiy*, "Subgraph Anomaly Detection in Financial Transaction Networks", in ACM, 2020, https://doi.org/10.1145/3383455.3422548

[8]. *Fabrizio Carcillo; Yann-Aël Le Borgne; Olivier Caelen; Yacine Kessaci; Frédéric Oblé; Gianluca Bontempi*, "Combining unsupervised and supervised learning in credit card fraud detection", Information Sciences, Volume 557, 2021, pp. 317-331, https://doi.org/10.1016/j.ins.2019.05.042.

[9]. *Chengliang Xu; Huanxin Chen*, "A hybrid data mining approach for anomaly detection and evaluation in res-idential buildings energy data", Energy and Buildings, Volume 215, 2020, 109864, https://doi.org/10.1016/j.enbuild.2020.109864

[10]. *Xue Liu; Yong Ding; Hao Tang; Feng Xiao*, "A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity consumption data", Energy and Buildings, Volume 231, 2021, 110601, https://doi.org/10.1016/j.enbuild.2020.110601

[11]. *Jose E. Cabral; Joao O. P. Pinto; Alexandra M. A. C. Pinto,* "Fraud detection system for high and low voltage electricity consumers based on data mining," 2009 IEEE Power & Energy Society General Meeting, Calgary, AB, Canada, 2009, pp. 1-5, doi: 10.1109/PES.2009.5275809

[12]. Bernat Coma-Puig; Josep Carmona; Ricard Gavaldà; Santiago Alcoverro; Victor Martin, "Fraud Detection in Energy Consumption: A Supervised Approach," 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Montreal, QC, Canada, 2016, pp. 120-129, doi: 10.1109/DSAA.2016.19.a

[13]. DELGAZ GRID - https://delgaz.ro/gaze-naturale – last access on January 20, 2023

[14]. E.ON Romania - https://www.eon.ro/ – last access on January 20, 2023

[15] WEKA - https://www.cs.waikato.ac.nz/ml/index.html – last access on January 20, 2023

[16]. *J. I. Guerrero, C. Leon, I. Monedero, F. Biscarri, and J. Biscarri.* Improving knowledge-based systems with sta-tistical techniques, text mining, and neural networks for non-technical loss detection. Knowledge-Based Sys-tems, 71:376 – 388, 2014, https://doi.org/10.1016/j.knosys.2014.08.014

[17]. *I. Monedero, F. Biscarri, C. Le´on, J. I. Guerrero, J. Biscarri, and R. Mill´an.* Detection of frauds and other non-technical losses in a power utility using pearson coefficient, bayesian

networks and decision trees. Interna-tional Journal of Electrical Power & Energy Systems, 34(1):90 – 98, 2012, https://doi.org/10.1016/j.ijepes.2011.09.009

[18]. *Abdelilah Ettaleby, Mohammed Boussetta, Mohamed Benslimane,* "Faults Detection for Photovoltaic Field Based on K-Means, Elbow, and Average Silhouette Techniques through the Segmentation of a Thermal Image", International Journal of Photoenergy, vol. 2020, Article ID 6617597, 7 pages, 2020. https://doi.org/10.1155/2020/6617597

[19]. *Alessandro Tallini, Luca Cedola,* "Evaluation Methodology for Energy Efficiency Measures in Industry and Service Sector", Energy Procedia, Vol. 101, 2016, pp. 542-549, https://doi.org/10.1016/j.egypro.2016.11.069.