IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 29, 2021, accepted May 24, 2021, date of publication June 2, 2021, date of current version June 10, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3085501

A Novel Features-Based Multivariate Gaussian Distribution Method for the Fraudulent Consumers Detection in the Power Utilities of Developing Countries

AMMAR YOUSAF KHARAL[1], HASSAN ABDULLAH KHALID[1], ADEL GASTLI[2], (Senior Member, IEEE), AND JOSEP M. GUERRERO[3], (Fellow, IEEE)

[1]U.S.-Pakistan Centers for Advanced Studies in Energy, National University of Science and Technology, Islamabad 44000, Pakistan
[2]Department of Electrical Engineering, Qatar University, Doha 2713, Qatar
[3]Department of Energy Technology, Aalborg University, 9100 Aalborg, Denmark

Corresponding author: Adel Gastli (adel.gastli@qu.edu.qa)

This work was supported by the Qatar National Library.

ABSTRACT According to statistics, developing countries all over the world have suffered significant non-technical losses (NTLs) both in natural gas and electricity distribution. NTLs are thought of as energy that is consumed but not billed e.g., theft, meter tampering, meter reversing, etc. The adaptation of smart metering technology has enabled much of the developed world to significantly reduce their NTLs. Also, the recent advancements in machine learning and data analytics have enabled a further reduction in these losses. However, these solutions are not directly applicable to developing countries because of their infrastructure and manual data collection. This paper proposes a tailored solution based on machine learning to mitigate NTLs in developing countries. The proposed method is based on a multivariate Gaussian distribution framework to identify fraudulent consumers. It integrates novel features like social class stratification and the weather profile of an area. Thus, achieving a significant improvement in fraudulent consumer detection. This study has been done on a real dataset of consumers provided by the local power distribution companies that have been cross-validated by onsite inspection. The obtained results successfully identify fraudulent consumers with a maximum success rate of 75%.

INDEX TERMS Artificial intelligence, data analytics, fraudulent consumer identification framework, machine learning, multivariate gaussian distribution, non-technical losses.

I. INTRODUCTION

Non-technical losses (NTLs) are considered as the energy that has flowed through the electricity and gas distribution networks to the end consumer but is not billed accordingly. These non-billing issues are usually caused by fraudulent activities such as theft, tampering of metering equipment, violating tariff obligations, etc. Both the developed and developing countries suffer from these NTLs. However, the remedies employed by the developed countries do not apply to developing countries due to scarce resources and cost constraints. One of the widely proposed solutions for NTLs reduction in developed countries is to upgrade the traditional grid and deploy smart meters [1]. However, due to the fragile economies of developing countries, such an upgrade would be a very costly endeavor. Furthermore, the developed countries are now using machine learning (ML) and data analytics (DA) techniques to detect NTLs. However, developing countries lack reliable systems for data collection and verification due to a lack of adequate resources and a transparent system [2]. Moreover, the reasons for fraud in electricity and gas consumption in developing and developed countries are also very different. Some prevalent reasons leading to fraud and its types in the electricity and gas distribution networks of the developing countries are presented in Table 1 [3], [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott.

TABLE 1. Factors leading to fraud in developing countries.

Sr. No.	Factors	Description
1	Poverty	Inability to pay utility bills
2	High Illiteracy	Lack moralities and awareness
3	Bad Governance	Failure of Institutions to deliver better service

These frauds committed on the power utilities cause financial losses of billions of dollars annually, whereas several billion dollars are spent on the detection and rectification of those frauds. Some of these illegal activities are also a risk to public safety (e.g. threat to life while tapping the live line) [5]. Most of the developing countries are already struggling with their economies, under such circumstances losses of power companies have a huge impact, particularly on circular debt and generally on the national exchequer [6]. Generally, the NTLs cause mammoth of financial losses all over the world, as stated in the World Bank's report 'over 50% of theft is in developing countries' [7]. NTLs faced by power utility companies of the United States were estimated between USD 1 to 10 billion [8]. The losses incurred in lieu of NTLs in India, Malaysia, Brazil, and the UK have amounted to USD 9 billion, 229 million, 5 billion, and £ 173 million annually, respectively [4], [8], [9]. While Pakistan witnessed 17.5% of these losses in FY 2012-13 and 16.9% in FY 2013-14 [10]. Similarly, theft of electricity in Turkey is reported to be 15.8% which is more than double the median of the OECD countries. In Spain, NTLs are estimated at 35%-45% [11]. In Canada, these losses are approximated at 100 million Canadian dollars every year, which if not stolen can supply electricity to 77,000 houses for a year [12]. Consequently, power utility companies lose above USD 25 billion each year worldwide (both developing and developed countries are included) [3]. The above NTLs are summarized in Figure 1 concerning the GDP of each country.

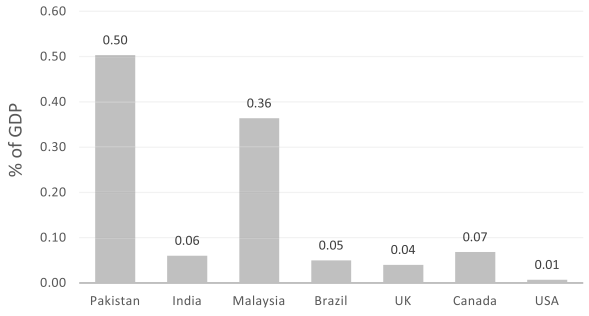


FIGURE 1. Non-technical losses as percentage of GDP of 2014.

Traditionally, there have been two broad categories of solutions to address the issue of NTLs, but each comes with its disadvantages. The first solution relies on excessive onsite inspections that include: door to door checking of each consumer by the power companies' inspection staff. Another solution is to install the latest infrastructure that includes enhanced metering facilities at the consumer end, distribution line monitoring, etc. There is a third category of solutions that have appeared with the development of artificial intelligence (AI) and data science (DS). Indeed in recent years, fraud detection and identification with the help of DA and ML has become the area of interest [4]. These data-oriented solutions are further categorized into supervised and unsupervised methods.

Numerous methods using supervised learning to detect fraud in electrical distribution systems include support vector machines are being used as binary classifiers and used in hybrid models for better classification. They require large datasets with response time ranging from months to days achieving hit-rates around 72% [13]–[17]. Most traditional artificial neural networks are used as a binary classifier for fraud detection. They also require large datasets with response time ranging from months to years [18]–[20]. Optimal path forest is being used in classification and clustering applications of fraud detection. Again, large datasets are required for training with response time ranging from months to years. Several other approaches have also been used in similar domains such as K-nearest neighbors algorithm (K-NN) classifiers [21], [22], condition-based rule induction methods [23], [24], and generalized additive models [25].

Whereas the unsupervised learning methods for detection and identification of NTLs in electrical distribution networks include: a special type of neural network known as self-organizing map appeared in literature for detection of NTLs. Producing visual representation of data that needed to be evaluated by the experts [26], [27]. Multiple types of clustering algorithms are used for fraud identification that enhances their classification and reduces false negative [22], [25], [28]. The system based on rules defined by the experts known as expert system played a critical role in surfacing fraudulent consumers from a pool of fraudulent and non-fraudulent consumers. They are usually dependent on utility inspectors to help accurately define rules [19]. Intermediate Monitor Meter (IMM) was proposed to analyze the power flow to detect the NTL with a detection accuracy of 95% [29]. Similarly, an innovative load-flow-based method that uses data of smart meters to determine NTLs provided promising results on unbalanced distribution systems [30]. Lately, a cost-effective and remote detection and identification method were proposed in [31], for detecting illegal electricity consumption while preserving the privacy of the consumers. An online evaluation method that leads to the replacement of faulty meters only by analyzing their power acquisition data is presented in [32].

In recent years, considerable improvement in fraudulent consumer detection has been reported for areas where large sets of labeled data are available. However, areas with small sets of unlabeled data are not much reported in the literature. Recently, an electricity theft detection (ETD) mechanism based on Relational Denoising Auto-encoder (RDAE) and Attention Guided (AG) TripleGAN is developed with the detection rate of 0.956 which makes it more acceptable than existing approaches [33]. A method based on Meter

Error Estimation is presented in [34], which adopts a decision tree to classify the abnormal data and resultantly highlights a malfunctional meter. Two-dimensional convolutional neural networks (CNN) and hybrid deep neural networks were employed to detect electricity theft in the smart grid. Huge datasets were used for learning and results were verified using the area under the curve. Fractional-order Self-Synchronization Error- Based Fuzzy Petri Nets were used to detect NTLs and outage scenarios. It has proved the practicality of methodology with the help of simulations on the IEEE 30 bus system [35]. Clustering-based novelty detection was employed to identify NTLs. It has achieved a true positive rate of 63.6%, a false positive rate of 24.3%, and obtained a 0.741 area under the curve (AUC) [36]. As many developing countries do not have such large, labeled datasets, thus, these methods cannot produce such results for their energy distribution companies. Due to the increasing shift from electronics meters to smart meters, there is an origination of novel countermeasures to the problem of NTL. Recently proposed methods include: extreme gradient boosted trees which use labeled data from a smart meter that has outperformed the rest of the classifiers by obtaining an AUC of 0.91 and a precision of 21% for on-field inspections [37]. In [38], the classifiers are combined with the Levenberg-Marquardt method to detect and identify illegal consumers in a smart grid environment. A binary black hole algorithm has been proposed in [39] that uses a metaheuristic optimization technique for theft characterization to minimize commercial losses in Brazil. A hierarchical model of smart grid networks and data collected from smart meters has been used in a generative model for anomaly detection in [40]. A comparison of 15 different ML techniques across nine types of classifiers is presented in [41]. Moreover, a feature selection framework is developed that highlighted fourteen features out of seventy one having a significant role in predicting NTLs. However, these recent ML-based solutions are not effective in developing countries because of the following reasons: Data collection by power utilities is extremely unreliable. Features included in the data collection of power companies are very limited. Power companies fear public consumers' consumption data. Consumption data is imbibed with political interference, bribery, social pressure, incompetence of staff, nepotism, etc. Utility staff is involved in manipulations of data to cover up losses. Fake reports are made to cover up fraudulent consumers.

In Pakistan, NTLs are never studied, because power utility companies lack research and development facilities and academia lack research and development resources and to date, there is no published evidence of research on NTLs in electricity and gas distribution utilities in Pakistan. This paper presents an ML-based framework to detect fraud in electricity and gas distribution utilities specifically for developing countries. This fraudulent consumer identification framework (FCIF) primarily uses the consumption data from electricity and gas distribution utilities and applies the unsupervised multivariate gaussian classifier (MGC) to separate the fraudulent consumers from non-fraudulent consumers. Before applying MGC, features are tailored to truly depict the behavior of fraudulent and non-fraudulent consumers. These features incorporate the socioeconomic and sociopolitical nature of Pakistani, Indian, and other developing societies as well as the composition of the particular human population and, thereby, select consumers for fraud mapping. Real data of 12752 consumers is used from the power utilities (both electricity and natural gas utilities) of Pakistan. This study focuses on detecting frauds, mentioned in Table 2, at the end of each month (since the billing in Pakistan is on monthly basis). The novelty of this research paper can be recapitulated as follows. This study focuses on the applicability of the solution to NTLs in the power sector of developing countries. Presented FCIF shows several merits. Big datasets are not needed. A novel feature that incorporates social class stratification and weather profile is introduced into the FCIF and it has significantly improved the results. A system with excessive computing power and memory is not required. As the threshold keeps moving until the hit (success) rate remains above 30%. The fraudsters are highlighted at the end of every month. An extensive onsite inspection is organized to verify the results produced by the FCIF achieving a hit rate of 75%, whereas the routine hit rate of power utilities is less than 5%. Resultantly, saves millions of dollars of power utility companies of Pakistan in lieu of NTLs, maintenance cost, onsite inspection cost, and additional monitoring staff cost, etc.

The rest of the paper is organized as follows. Section II presents the related literature of Multivariate Gaussian Distribution. FCIF model and detailed methodology are presented and discussed in section III. The cross-validated results are shown and discussed in Section IV. Finally, the conclusion is given in section V.

TABLE 2. Causes of NTLs in developing countries.

Sr. No.	Factors	Description
1	By-passing meter	Directly tapping supply lines to prevent meter from registering energy consumption
2	Meter tampering	Obstructing meter from accurately registering energy consumption
3	Meter reversal	Reversing the registered consumption to lower the consumption bill
4	Multiple meters of same utility on single property	Intended to stay in the lower tariff slab by distributing consumption
5	Change of tariff	Commercial use from domestic meters to avoid high tariff billing
6	Extension of electricity network and gas house line	Supplying electricity and gas to neighbors without the consent of Energy Utility
7	Use of compressors in houses (to get more gas from gas network)	Can damage the gas supply line and also affect the gas supply to others in the vicinity
8	Electricity generation using natural gas	Electricity generation on domestic gas tariff is a violation of tariff, secondly most of the gas generators contain in built compressor which is again a violation.

II. MULTIVARIATE GAUSSIAN DISTRIBUTION

This section discusses the essentials of Multivariate Normal (or Gaussian) Distribution (MGD). Detailed literature, mathematics certain observations, and different propositions can be found in [42]. The functionality of MGD is that it constructs a probability distribution bell-shaped curve centered at the mean (μ). The population having large variance (σ) and divergent behavior from the majority population lies far away from the mean, while the population with low variance and similar behavior to others in the population lies closer to the mean. Thus, the population closer to the mean has higher probabilities, whereas the population farther from mean has lower probabilities. The lower probability population is considered as anomalous as compared to the majority population. Additional rules can be incorporated to refine the anomalous population.

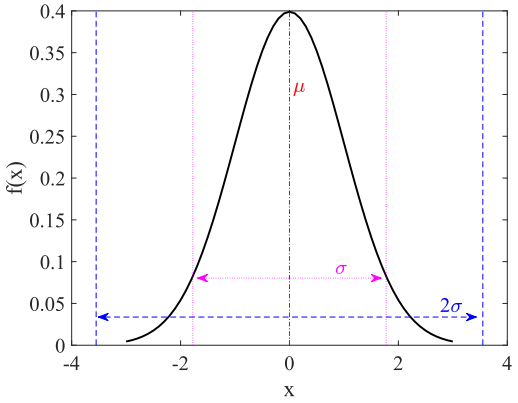


FIGURE 2. Univariate gaussian distribution.

MGD is the multivariate generalization of the univariate normal distribution as in Figure 2 and expressed as $X \sim N\left(\mu, \sum\right)$ and joint probability density function of X is given as

$$p(X; \mu, \sum) = \frac{e^{-\frac{1}{2}(X-\mu)^T \sum^{-1}(X-\mu)}}{\sqrt{(2\pi^n)|\sum|}} \tag{1}$$

Here $X = x_1, x_2, x_3 \ldots x_n$ is the $n-$dimensional vector where $x_1, x_2, x_3 \ldots x_n$ are $(m \times 1)$ dimensional column vectors. μ is a $(1 \times n)$ dimensional vector containing mean of the column vectors $x_1, x_2, x_3 \ldots x_n$. $\sum$ is the $(n \times n)$ dimensional covariance matrix. The exponent of e contains product of the transpose of $(X - \mu)$, inverse of $\sum$ and $(X - \mu)$, its dimension is $(1 \times n)\,(n \times n)\,(n \times 1) = (1 \times 1)$. Note that the term $\frac{1}{\sqrt{(2\pi^n)|\sum|}}$ contains only constants and mainly act as a normalization factor. The most significant term of the probability distribution function is covariance; $\sum = cov\left(x_i, x_j \ldots x_n\right)$. It results in $(n \times n)$ dimensional matrix which implies the interdependence of features n.

A. VARIANCE-COVARIANCE MATRIX

The covariance matrix, also known as autocovariance matrix, as, mentioned in (2).

$$\sum = \begin{bmatrix} a_{ii} & a_{ij} & \cdots & a_{in} \\ a_{ji} & ajj & \ldots & a_{jn} \\ \vdots & \vdots & \ddots & \vdots \\ a_{ni} & a_{nj} & \cdots & a_{nn} \end{bmatrix} \tag{2}$$

where the main diagonal entries $\left(a_{ii}, a_{jj} \ldots a_{nn}\right)$ illustrate variance within the column vectors $x_i, x_j \ldots x_n$, whereas, non-diagonal entries of covariance matrix express the covariance between i^{th} and j^{th} feature of random vector $X = x_i, x_j \ldots x_n$. If non-diagonal entries are zero, it signifies that column vectors $x_i, x_j \ldots x_n$ are uncorrelated and independent of each other.

B. MAHALANOBIS DISTANCE

The expression in the exponent of e, $(X - \mu)^T \left(\sum\right)^{-1} (X - \mu)$ is known as squared **Mahalanobis distance** between X and μ. This gives the $n-$dimensional bell-shaped curve. In which every element of vector X is positioned, concerning its Mahalanobis distance, from mean μ. The highest probabilities are of those values of X which are placed nearer to the multivariate mean vector μ and represents the common pattern of major population of vector X. The values positioned farther from multivariate mean vector μ have lower probabilities and are thus considered inconsistent as compared to the other population of vector X.

III. METHODOLOGY AND DISCUSSION

This section presents a detailed and sequential step to detect fraud in electricity and natural gas consumption while separating the fraudulent consumers from non-fraudulent consumers. The proposed framework to classify fraudulent energy customers from non-fraudulent energy customers is presented in Figure 3 and the general representation of the system under study is shown in Figure 4.

A. DATA SETS

This study uses monthly consumption datasets for electricity and natural gas from the Lahore Electricity Supply Company (LESCO) and Sui Northern Gas Pipelines Limited (SNGPL) from which some sample data is presented in Figure. 8(b) and (d), respectively. The data set consists of domestic consumers of both electricity (3255) and natural gas (9496), containing different features as compared in Table 3. It is pertinent to mention here that data collection of power companies in Pakistan is not very extensive as very limited features are available for analysis. Moreover, the initial request for labeled data was denied due to privacy concerns. Furthermore, the utilities stated that by undertaking a large exercises of labeling the fraudulent consumers may result in cautious behavior of the fraudulent consumer to avoid the process, as it happened during the inspection process.

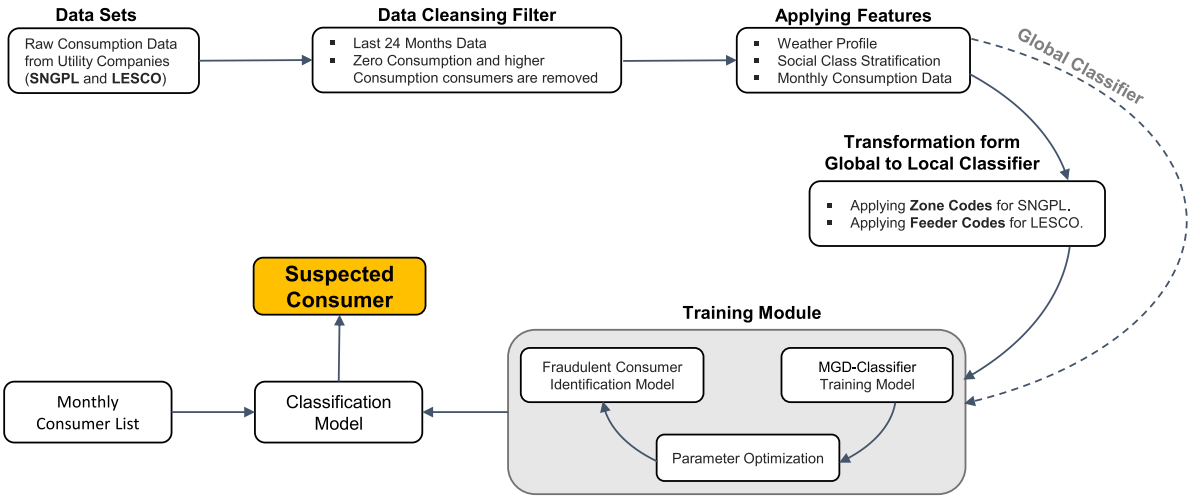


FIGURE 3. Fraudulent consumer identification framework.

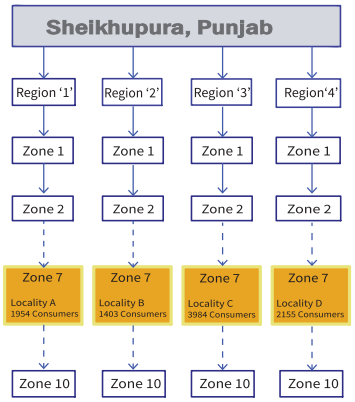


FIGURE 4. General representation of localities in this study.

TABLE 3. Comparison of available features.

Sr. No.	LESCO	SNGPL
1	**Units to be billed:** Electricity consumed by consumer in kilowatt hour (kWh).	**Volume of gas to be billed:** Consumption of Gas by consumer in hecta meter cube (Hm^3).
2	**Feeder Code:** Region is divided into sub-regions using feeders.	**Zone Code:** Region is divided into several sub-regions using zone codes.
3	**Payment history:** All past bills issued and their payment record.	**Payment history:** All past bills issued and their payment record.
4	**Units adjusted:** Units added or subtracted from monthly consumption on various grounds (penalty, overbilling etc.)	**Units adjusted:** Units added or subtracted from monthly consumption on various grounds (penalty, overbilling etc.)
5	**Sanctioned load:** in kilowatts (kW) allowed by the utility for a particular consumer.	

Therefore, the utility companies request the surprise visit. This shortcoming has raised many challenges and limited our choice to very few unsupervised classifiers for fraud detection. The real challenge which comes after data acquisition is the cleansing of data before the extraction of features.

B. DATA CLEANSING

Real-world data contains a lot of noise which disrupts the pattern hidden inside the data. A filter is designed to clean the data so that meaningful fraud patterns could be mapped. By analyzing the data with the help of experts of LESCO and SNGPL following parameters of data cleansing filter are selected:

- Past 24 month's consumption data is selected. The more we go into the past more the consumption pattern gets noisy because either old resident shifted out or new residents settled in the region, similarly, tariff rates were also lower in the past.
- Consumers with zero units to be billed in a particular month in two years are not included in the data set from which features are extracted. Such consumers adversely affect the average consumption pattern of the region.
- Consumers, whose meters are installed after January 2017, are also removed from the data set of feature extraction.
- Consumers with continuously large units to be billed throughout the year are also removed from the data set of feature extraction.

After applying the data cleansing filter on raw data, a relatively less noisy data set is attained. Now the next challenging stage is feature extraction.

C. FEATURE IDENTIFICATION AND SELECTION

In developing countries, like Pakistan, the data which is available with power companies are also subjected to political influences, peer pressures and manipulations by utility engineers to adjust the losses in the monthly billing of consumers. We are, therefore, constrained to only one feature i.e. consumption. Other features are selected in a way that could help us in selecting the consumption data in such a way that could map the effective fraud pattern. The other features are:

· **Weather profile** (humidity, temperature, UV intensity).

· **Social status of consumers** (Elite class, middle-class, lower-middle-class, below poverty line class).

It is found that in developing countries social class stratification is very prevalent. Usually, the people living in the same vicinity belong to the same class with few exceptions. Therefore, using consumption data of several regions collectively does not produce realistic fraud mapping. Therefore, consumption data is divided into several regions using the zone code of SNGPL and feeder code of LESCO, while additional help is also sought from billing book numbers. Another important thing that needs to be noticed is that after the data is divided into different regions, the weather profile also becomes constant in that region. A comparison of weather profiles of Sheikhupura and Islamabad can be seen in Figure 5 [43]. While the social class stratification for comparison of a consumer from a rich and poor neighborhoods can be visualized in Figure 6.

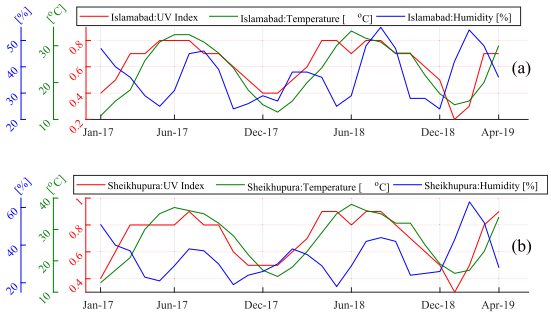


FIGURE 5. Comparison of UV intensity, temperature and humidity of (a) Islamabad and (b) Sheikhupura.

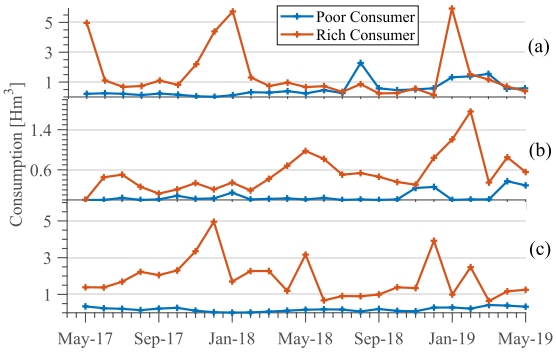


FIGURE 6. Consumption comparison of two social status consumers for Zones (a) 6830, (b) 6834, and (c) 6931.

Furthermore, for example, *if the fraud is needed to be detected for the month of December 2018, then the data set will be comprised of consumption data for the month of December 2017 and December 2018 of all the consumers in the selected locality*. After applying the respective data cleansing filter and selecting the above-mentioned features, scatter plots of the data set of SNGPL and LESCO are visualized as in Figure 7.

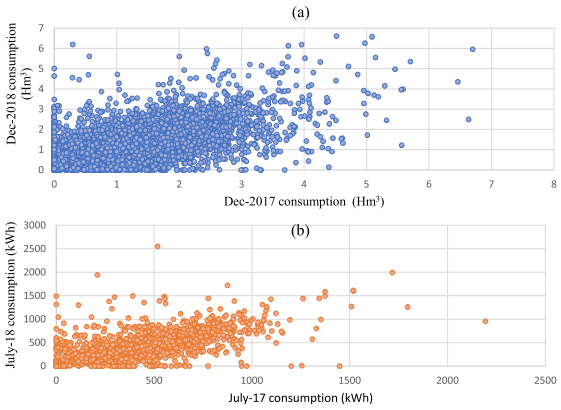


FIGURE 7. Scatter plot of filtered data from (a) SNGPL and (b) LESCO.

D. SELECTION OF A CLASSIFIER

Limited data with few features have limited our option to few classifiers. For the classification of this data, MGD is used and the following steps have led us to its selection.

· Data is unlabeled and there is no information who is the fraudulent consumer and who is not. Hence, an unsupervised classifier is required to isolate the fraudster.

· Data sets are small (data of conventional meters' monthly reading since smart meters are not available).

· Raw data is not very rich and, consequently, features are very limited.

· A simpler classifier with a flexible threshold setting is required so that employees of power companies could incorporate it into their routine work.

· In our case, MGD has produced results that are either better or equivalent than the state-of-the-art classifiers.

MGD returns the probabilities of different consumers of being fraudulent. Post-processing is performed on the data after getting their probabilities of being fraudulent, which leads to the final list of suspected customers.

E. THRESHOLD SELECTION

In Gaussian distribution curves, three-standard-deviations are accepted norm for the threshold of anomaly detection. However, in our case, it is not acceptable. After consulting relevant authorities of LESCO and SNGPL and depending upon the availability of resources, the threshold was adjusted to shortlist 1 to 5 % of most suspected consumers for onsite physical inspection.

F. FRAUDS DETECTED

Firstly, FCIF is suitable and workable for detecting following frauds in SNGPL:

· **Change of tariff (commercial use from domestic tariff):** Gas consumption of fraudsters is increased from the regular consumption pattern of domestic consumers.

· **Meter tampering:** Gas consumption of the fraudster is decreased from regular domestic consumption patterns.

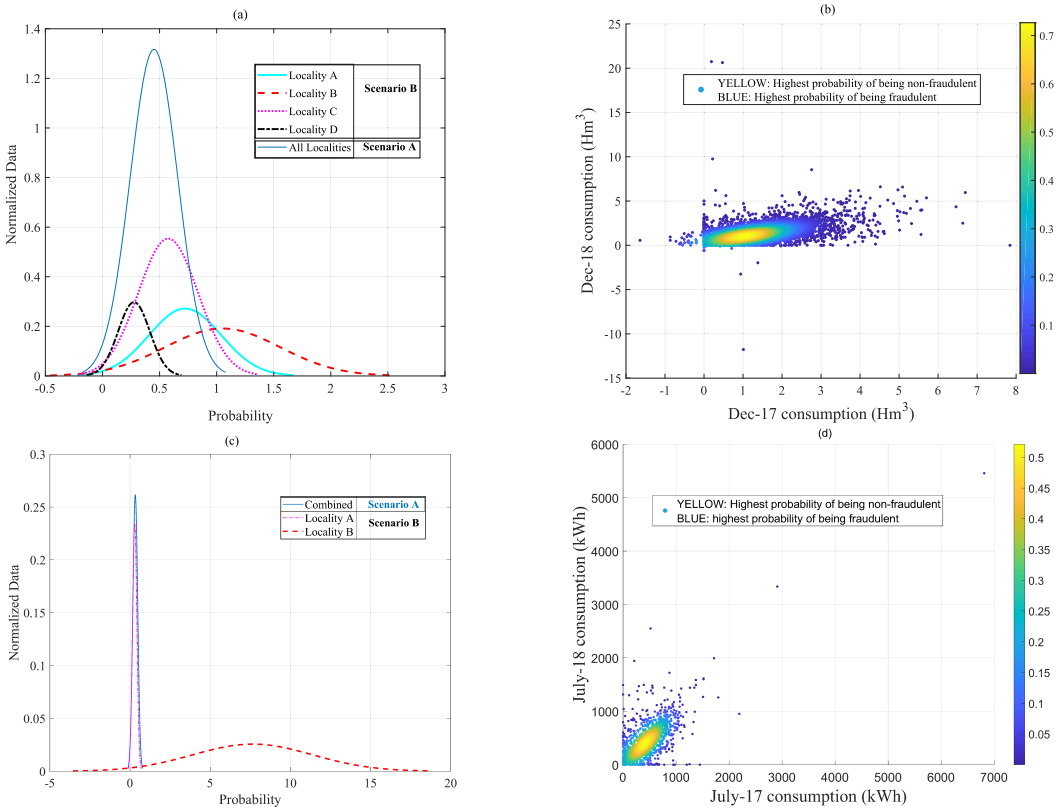


FIGURE 8. Scenario A: SNGPL Data (a) Probability distribution and (b) consumption data heat map, LESCO data (c) Probability distribution and (d) consumption data heat map.

· **By-passing meter:** Gas consumption of fraudster is decreased sharply as compared to the regular domestic consumption.

· **Illegal extension of house line:** In this case, the consumption of the defaulter is increased to a large extent than the regular domestic consumer.

· **Reversal of meter:** In this case, the consumption is either dramatically decreased or has an abrupt consumption pattern.

· **Illegal use of compressor:** In this case, the consumption is greater than the neighboring consumers.

· **Electricity generation from natural gas:** In this case, the consumption is greater than other domestic gas consumers in the vicinity.

Secondly, FCIF is suitable and workable for detecting following frauds in LESCO:

· **Meter tampering:** Electricity consumption is asymmetrical and is much lower than the other consumers in the vicinity.

· **By-passing meter:** Electricity consumption is much lower than that of other consumers in the vicinity.

· **Reversal of meter:** In this case, the consumption is either dramatically decreased or has an abrupt consumption pattern.

· **Change of tariff (commercial use from domestic tariff):** Consumption of fraudsters is increased from the regular consumption pattern of domestic consumers.

· **Illegal extension of house wiring:** Electricity consumption is glaringly high as compared to other consumers in the vicinity.

IV. VALIDATION OF RESULTS AND DISCUSSION

To prove the authenticity and utility of the new feature (i.e. social class stratification) and applicability of FCIF in developing countries, the following two scenarios have been considered, which demonstrate the improved results obtained by using the newly designed feature.

A. SCENARIO A: WITHOUT SOCIOECONOMIC AND WEATHER PROFILES

In scenario A, the novel feature which incorporates the socioeconomic nature of Pakistani society and weather profile is ignored. Consequently, the data of all localities get combined into one data set and is then passed on to the FCIF. The classifier in this scenario act as global learner classifier where all the data is used collectively to train a general classifier, as presented in Figure 3. The resultant probability distribution and its heat map are shown in Figure 8. It can be seen in the heat map that the zone with yellowish color shows the highest density of consumers with more or less similar consumption, hence the lowest probability of being fraudulent. As the variance in consumption increases color

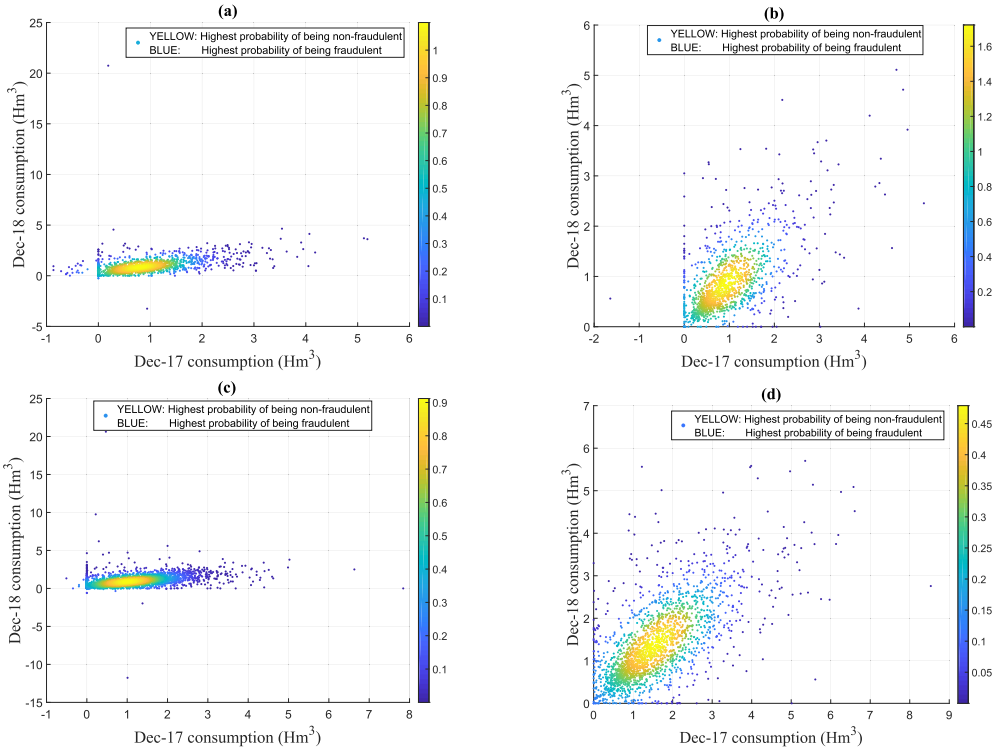


FIGURE 9. SNGPL Data for Scenario B: Consumption data heat map of (a) locality A, (b) locality B, (c) locality C and (d) locality D.

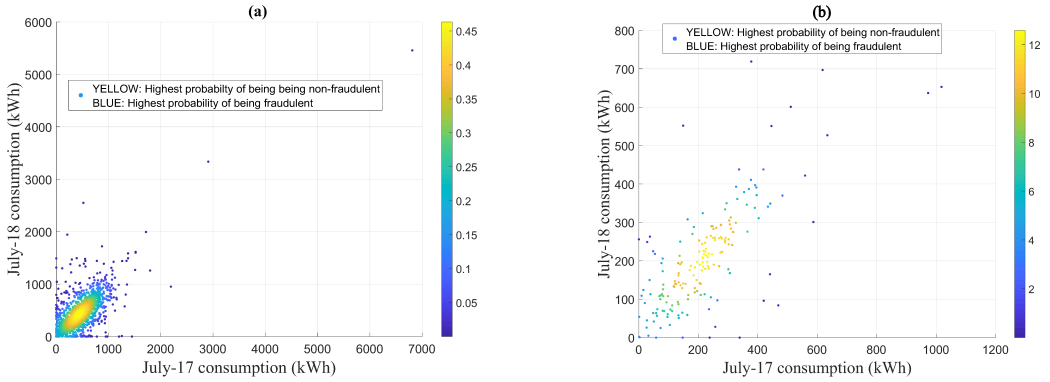


FIGURE 10. LESCO Data for Scenario B: Consumption data heat map of (a) locality A and (b) locality B.

converts into sky blue to blue, where blue color signifies the highest probability of being fraudulent.

B. SCENARIO B: WITH SOCIOECONOMIC AND WEATHER PROFILES

In scenario B, the features which employ socioeconomic and weather profile are included and, resultantly, the SNGPL and LESCO data set are transformed into datasets (i.e. A, B, C, D, where the demography is consistent). Now, instead of one general classifier we have got as much classifiers as there are zone codes. Zone codes are very carefully labeled on to the consumers taking into consideration the geospatial features, as presented in Figure 3. Now, once the classifier moved from Global to Local our classifiers get independence from the covariate shift impact [44].

These datasets are then passed to the FCIF. The resultant probability distribution of locality A, B, C, D are presented in Figure 8(a) and (c), while the consumption data's heat map are shown in Figures 9 and 10. We can see in all the heat maps that once the data is divided into several data sets the pattern of consumption comes out to be different for each locality. Resultantly, all the fraudsters are more visible and easily detected by the FCIF.

C. COMPARISON OF SCENARIO A AND SCENARIO B

The new feature that signifies the social class stratification of the society and weather profile, in the developing world, plays a very significant role in improving fraud detection results. For instance, the energy consumption of people living in slums cannot be compared with the energy consumption of people living in modern luxurious housing. Almost all the papers recently published and also cited in this paper use the aggregated data from the power companies. The actual reason for the decrease in the detection rate in scenario A is that when FCIF is applied on aggregated data incorporating all neighborhoods, the anomalous consumers get buried inside the variance of the accumulated data. For instance, the highest consumption of a family living in slums is lower than any of the family living in luxurious housing. Hence when combined together, the lower consumption frauds of luxurious housing fall within the cluster of slum consumption and, consequently, cannot be detected. Whereas when the FCIF analyses the data after incorporating the class stratification feature, as in scenario B, such consumers get highlighted easily. Hence the detection rate increases as in scenario B. The summarized results of both experiments are presented in Table 4.

TABLE 4. Comparison of scenarios A & B.

		Locality	No of Consumers	Detection Threshold %	Fraud Consumers	Hit-rate %
SNGPL	Scenario B	A	1954	5	98	71
		B	1403	5	71	69
		C	3984	5	200	70
		D	2155	5	108	73
	Scenario A	Combined	9496	5	475	43
LESCO	Scenario B	A	2064	1	21	75
		B	1191	1	12	70
	Scenario A	Combined	3255	1	33	44

D. ONSITE CROSS-VALIDATION

Results obtained from the proposed FCIF were sent to the respective power utilities. The feedback from SNGPL, the largest Natural Gas Distribution Company of Pakistan, and LESCO, Lahore Electric Supply Company, have been received. Cross verification of consumers, highlighted by the framework presented in this paper, through onsite inspection by the energy company's inspection staff reflects the effectiveness of the proposed framework. Summarized results are displayed in Table 4, with a comparison of other latest techniques and their results are displayed in the Table 5, however different methods and data sets have been used in these algorithms.

TABLE 5. Comparison with recent NTLs detection methods.

Algorithm	Data set	Labeled Data	Metrics	Results	Ref
SVM-FIS	Large	-	Hit rate	72%	[23]
Clustering-based novelty detection	Large	Yes	TP FP AUC	63.6% 24.3% 0.741	[36]
Binary Black Hole algorithm	Large	Yes	Mean accuracy rates	64.81%	[39]
Convolution Neural Networks	Large	Yes	AUC MAP	0.8001 0.9565	[45]
DT-KSVM	Small	Yes	AUC	0.956	[46]
MOD-WPT	Small	Yes	AUC Recall Accuracy	0.8187 0.65 0.9435	[47]
Light Gradient Boosting and Adapative Boosting	Med	Yes	Precision Recall MCC F1 Score	0.968 0.94 0.91 0.95	[48]
Proposed Method	Small	No	Hit rate	75%	-

Verified results prove that the highest hit rate is 73% in locality D for SNGPL, whereas for LESCO it is 75%. In locality D, the 2155 consumers' consumption data is used and after applying a threshold of 5%, 108 consumers were detected as fraudsters. However, when onsite cross-validation was done 73% were found real fraudsters. Similarly, for LESCO threshold was dropped to 1% due to onsite inspection staff constraints and hit-rate came out to be 75%.

The detailed analysis of highlighted fraudster consumers shows that certain types of frauds (e.g. commercial use, illegal house line extension, and meter reversing) are being committed in the locality D, where as live line tapping, meter tampering and multiple meters on single premises are more common frauds in locality A of LESCO. Also it was observed that all localities do not hold the same type of frauds. It mostly depends on the social class stratification of a particular locality which is predominant in the developing world.

V. CONCLUSION

In this paper, the fraudulent consumer identification framework to mitigate non-technical losses in natural gas and electricity distribution companies is validated. This framework is presented while acknowledging the shortcomings and limitations faced by developing countries in mitigating frauds in energy consumption. Multivariate Gaussian distribution is employed in FCIF wherein features like social class stratification and weather variations further facilitated MGD in the realistic mapping of fraud. Obtained results are validated by onsite inspection taken out by energy distribution companies on the prediction results of FCIF. Crosschecking results have demonstrated that the proposed framework has a maximum hit-rate of 75%. Thus, outperformed all other published frameworks as they have not incorporated the peculiarity of developing countries.

REFERENCES

[1] C. Gellings, "Estimating the costs and benefits of the smart grid: A preliminary estimate of the investment requirements and the resultant benefits of a fully functioning smart grid," Electr. Power Res. Inst., Washington, DC, USA, Tech. Rep. 1022519, 2011, vol. 1.

[2] A. Elahi, "Challenges of data collection in developing countries—The Pakistini experience as a way forward," *Stat. J. IAOS, J. Int. Assoc. Off. Statist.*, vol. 25, nos. 1–2, pp. 11–17, 2008. [Online]. Available: https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji00681

[3] S. S. S. R. Depuru, L. Wang, and V. Devabhaktuni, "Electricity theft: Overview, issues, prevention and a smart meter based approach to control theft," *Energy Policy*, vol. 39, no. 2, pp. 1007–1015, Feb. 2011.

[4] J. L. Viegas, P. R. Esteves, R. Melício, V. M. F. Mendes, and S. M. Vieira, "Solutions for detection of non-technical losses in the electricity grid: A review," *Renew. Sustain. Energy Rev.*, vol. 80, pp. 1256–1268, Dec. 2017.

[5] A. J. Taylor, G. McGwin, R. M. Brissie, L. W. Rue, and G. G. Davis, "Death during theft from electric utilities," *Amer. J. Forensic Med. Pathol.*, vol. 24, no. 2, pp. 173–176, 2003.

[6] P. Glauner, C. Glaeser, N. Dahringer, P. Valtchev, R. State, and D. Duarte, "Non-technical losses in the 21st century: Causes, economic effects, detection and perspectives," Univ. Luxembourg, Luxembourg City, Luxembourg, 2018. [Online]. Available: https://www.glauner.info/publications

[7] P. Antmann, "Reducing technical and non-technical losses in the power sector," World Bank, Washington, DC, USA, 2009. [Online]. Available: https://openknowledge.worldbank.org/handle/10986/20786

[8] T. B. Smith, "Electricity theft: A comparative analysis," *Energy Policy*, vol. 32, no. 18, pp. 2067–2076, Dec. 2004.

[9] A. Lazaropoulos, "Detection of energy theft in overhead low-voltage power grids—The hook style energy theft in the smart grid era," *Trends Renew. Energy*, vol. 5, no. 1, pp. 12–46, Feb. 2019.

[10] T. Ahmad, "Non-technical loss analysis and prevention using smart meters," *Renew. Sustain. Energy Rev.*, vol. 72, pp. 573–589, May 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364032117300990

[11] I. Monedero, F. Biscarri, C. León, J. I. Guerrero, J. Biscarri, and R. Millán, "Detection of frauds and other non-technical losses in a power utility using pearson coefficient, Bayesian networks and decision trees," *Int. J. Electr. Power Energy Syst.*, vol. 34, no. 1, pp. 90–98, Jan. 2012.

[12] BCHydro. (2011). *Smart Meters Help Reduce Electricity Theft, Increase Safety*. [Online]. Available: https://www.bchydro.com/news/conservation/2011/smart_meters_energy_theft.html

[13] M. Di Martino, F. Decia, J. Molinelli, and A. Fernández, "A novel framework for nontechnical losses detection in electricity companies," in *Pattern Recognition-Applications and Methods*. Berlin, Germany: Springer, 2013, pp. 109–120.

[14] Y. Guo, C.-W. Ten, and P. Jirutitijaroen, "Online data validation for distribution operations against cybertampering," *IEEE Trans. Power Syst.*, vol. 29, no. 2, pp. 550–560, Mar. 2014.

[15] G. M. Messinis and N. D. Hatziargyriou, "Review of non-technical loss detection methods," *Electr. Power Syst. Res.*, vol. 158, pp. 250–266, May 2018.

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

[17] A. Jindal, A. Dua, K. Kaur, M. Singh, N. Kumar, and S. Mishra, "Decision tree and SVM-based data analytics for theft detection in smart grid," *IEEE Trans. Ind. Informat.*, vol. 12, no. 3, pp. 1005–1016, Jun. 2016.

[18] A. H. Nizar, Z. Y. Dong, and Y. Wang, "Power utility nontechnical loss analysis with extreme learning machine method," *IEEE Trans. Power Syst.*, vol. 23, no. 3, pp. 946–955, Aug. 2008.

[19] J. I. Guerrero, C. León, I. Monedero, F. Biscarri, and J. Biscarri, "Improving knowledge-based systems with statistical techniques, text mining, and neural networks for non-technical loss detection," *Knowl.-Based Syst.*, vol. 71, pp. 376–388, Nov. 2014.

[20] B. C. Costa, B. L. Alberto, A. M. Portela, W. Maduro, and E. O. Eler, "Fraud detection in electric power distribution networks using an ann-based knowledge-discovery process," *Int. J. Artif. Intell. Appl.*, vol. 4, no. 6, p. 17, 2013.

[21] C. C. O. Ramos, A. N. de Souza, A. X. Falcao, and J. P. Papa, "New insights on nontechnical losses characterization through evolutionary-based feature selection," *IEEE Trans. Power Del.*, vol. 27, no. 1, pp. 140–146, Jan. 2012.

[22] S. Y. Han, J. No, J. Shin, and Y. Joo, "Conditional abnormality detection based on AMI data mining," *IET Gener., Transmiss. Distrib.*, vol. 10, no. 12, pp. 3010–3016, Sep. 2016.

[23] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and F. Nagi, "Improving SVM-based nontechnical loss detection in power utility using the fuzzy inference system," *IEEE Trans. Power Del.*, vol. 26, no. 2, pp. 1284–1285, Apr. 2011.

[24] C. León, F. Biscarri, I. Monedero, J. I. Guerrero, J. Biscarri, and R. Millán, "Variability and trend-based generalized rule induction model to NTL detection in power companies," *IEEE Trans. Power Syst.*, vol. 26, no. 4, pp. 1798–1807, Nov. 2011.

[25] L. T. Faria, J. D. Melo, and A. Padilha-Feltrin, "Spatial-temporal estimation for nontechnical losses," *IEEE Trans. Power Del.*, vol. 31, no. 1, pp. 362–369, Feb. 2016.

[26] C. C. O. Ramos, A. N. Souza, G. Chiachia, A. X. Falcão, and J. P. Papa, "A novel algorithm for feature selection using harmony search and its application for non-technical losses detection," *Comput. Electr. Eng.*, vol. 37, no. 6, pp. 886–894, Nov. 2011.

[27] C. C. O. Ramos, A. N. de Sousa, J. P. Papa, and A. X. Falcao, "A new approach for nontechnical losses detection based on optimum-path forest," *IEEE Trans. Power Syst.*, vol. 26, no. 1, pp. 181–189, Feb. 2011.

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

[29] T. S. D. Ferreira, F. C. L. Trindade, and J. C. M. Vieira, "Load flow-based method for nontechnical electrical loss detection and location in distribution systems using smart meters," *IEEE Trans. Power Syst.*, vol. 35, no. 5, pp. 3671–3681, Sep. 2020.

[30] A. Bin-Halabi, A. Nouh, and M. Abouelela, "Remote detection and identification of illegal consumers in power grids," *IEEE Access*, vol. 7, pp. 71529–71540, 2019.

[31] H. Cai, H. Chen, X. Ye, X. Zhang, H. Wen, J. Li, and Q. Guo, "An online state evaluation method of smart meters based on information fusion," *IEEE Access*, vol. 7, pp. 163665–163676, 2019.

[32] Z. Aslam, F. Ahmed, A. Almogren, M. Shafiq, M. Zuair, and N. Javaid, "An attention guided semi-supervised learning mechanism to detect electricity frauds in the distribution systems," *IEEE Access*, vol. 8, pp. 221767–221782, 2020.

[33] F. Liu, C. Liang, and Q. He, "Remote malfunctional smart meter detection in edge computing environment," *IEEE Access*, vol. 8, pp. 67436–67443, 2020.

[34] S.-J. Chen, T.-S. Zhan, C.-H. Huang, J.-L. Chen, and C.-H. Lin, "Nontechnical loss and outage detection using fractional-order self-synchronization error-based fuzzy Petri nets in micro-distribution systems," *IEEE Trans. Smart Grid*, vol. 6, no. 1, pp. 411–420, Jan. 2015.

[35] J. Y. Kim, Y. M. Hwang, Y. G. Sun, I. Sim, D. I. Kim, and X. Wang, "Detection for non-technical loss by smart energy theft with intermediate monitor meter in smart grid," *IEEE Access*, vol. 7, pp. 129043–129053, 2019.

[36] J. L. Viegas, P. R. Esteves, and S. M. Vieira, "Clustering-based novelty detection for identification of non-technical losses," *Int. J. Electr. Power Energy Syst.*, vol. 101, pp. 301–310, Oct. 2018.

[37] M. M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero, and A. Gómez-Expósito, "Detection of non-technical losses using smart meter data and supervised learning," *IEEE Trans. Smart Grid*, vol. 10, no. 3, pp. 2661–2670, May 2019.

[38] A. A. Ghasemi and M. Gitizadeh, "Detection of illegal consumers using pattern classification approach combined with Levenberg-Marquardt method in smart grid," *Int. J. Electr. Power Energy Syst.*, vol. 99, pp. 363–375, Jul. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0142061517314448

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

[40] R. Moghaddass and J. Wang, "A hierarchical framework for smart grid anomaly detection using large-scale smart meter data," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 5820–5830, Nov. 2018.

[41] K. M. Ghori, R. A. Abbasi, M. Awais, M. Imran, A. Ullah, and L. Szathmary, "Performance analysis of different types of machine learning classifiers for non-technical loss detection," *IEEE Access*, vol. 8, pp. 16033–16048, 2020.

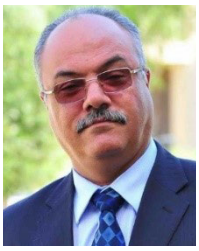
[42] C. B. Do and H. Lee, "Section notes 9—Gaussian processes," Lect. Notes, 2008, pp. 1–14. [Online]. Available: https://scholar.google.com/scholar_lookup?title=Section%20notes%209%20-%20Gaussian%20processes&publication_year=2008&author=C.B.%20Do&author=H.%20Lee
[43] *Weather Data for Sheikhupura and Islamabad*. [Online]. Available: https://www.worldweatheronline.com/islamabad-weather-averages/islamabad/pk.aspx
[44] P. Glauner, "Artificial intelligence for the detection of electricity theft and irregular power usage in emerging markets," Ph.D. dissertation, Univ. Luxembourg, Luxembourg City, Luxembourg, 2019.
[45] Z. Zheng, Y. Yang, X. Niu, H.-N. Dai, and Y. Zhou, "Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids," *IEEE Trans. Ind. Informat.*, vol. 14, no. 4, pp. 1606–1615, Apr. 2018.
[46] X. Kong, X. Zhao, C. Liu, Q. Li, D. Dong, and Y. Li, "Electricity theft detection in low-voltage stations based on similarity measure and DT-KSVM," *Int. J. Electr. Power Energy Syst.*, vol. 125, Feb. 2021, Art. no. 106544. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S014206152030702X
[47] N. F. Avila, G. Figueroa, and C.-C. Chu, "NTL detection in electric distribution systems using the maximal overlap discrete wavelet-packet transform and random undersampling boosting," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 7171–7180, Nov. 2018.
[48] A. Aldegheishem, M. Anwar, N. Javaid, N. Alrajeh, M. Shafiq, and H. Ahmed, "Towards sustainable energy efficiency with intelligent electricity theft detection in smart grids emphasising enhanced neural networks," *IEEE Access*, vol. 9, pp. 25036–25061, 2021.

AMMAR YOUSAF KHARAL received the B.E. degree in electrical engineering from the School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Islamabad, Pakistan, in 2014, where he is currently pursuing the M.S. degree in electrical engineering (power) with the U.S.-Pakistan Centers for Advanced Studies in Energy.

Since August 2015, he has been working as a Distribution Engineer with Sui Northern Gas Pipelines Ltd., Punjab, Pakistan, where he worked in operations and maintenance of distribution network, UFG control, and metering. His research interests include machine learning, artificial intelligence, data science, power system optimization, renewable energy, and energy conservation and efficiency.

HASSAN ABDULLAH KHALID received the B.Sc. degree in electrical engineering from Air University, Islamabad, Pakistan, in 2007, the M.Sc. degree in electrical power engineering from the Chalmers University of Technology, Gothenburg, Sweden, in 2010, and the Ph.D. degree from the University of L'Aquila, L'Aquila, Italy, in 2016.

Since 2016, he has been with the National University of Science and Technology, Islamabad, where he is currently an Assistant Professor. His current research interests include systems control with applications in power electronics, energy conversion, renewable energy, and smart grids.



ADEL GASTLI (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from the National School of Engineers of Tunis, Tunisia, in 1985, and the M.Sc. and Ph.D. degrees in electrical and computer engineering from the Nagoya Institute of Technology, Japan, in March 1990 and March 1993, respectively.

From September 1985 to September 1987, he worked with the National Institute for Standards and Intellectual Property, Tunisia. He worked with Mitsubishi Electric Corporation, Japan, from April 1993 to July 1995. He joined the Electrical and Computer Engineering Department, Sultan Qaboos University, Oman, in August 1995. He was the Head of the Department, from September 2001 to August 2003 and from September 2007 to August 2009. He was appointed as the Director of Sultan Qaboos University Quality Assurance Office, from February 2010 to January 2013. In February 2013, he joined the Electrical Engineering Department, Qatar University, as a Professor, and the Kahramaa-Siemens Chair of Energy Efficiency. From August 2013 to September 2015, he was appointed as the Associate Dean of Academic Affairs at the College of Engineering. His current research interests include energy efficiency, renewable energy, electric vehicles, and smart grid.

JOSEP M. GUERRERO (Fellow, IEEE) received the B.S. degree in telecommunications engineering, the M.S. degree in electronics engineering, and the Ph.D. degree in power electronics from the Technical University of Catalonia, Barcelona, in 1997, 2000, and 2003, respectively.

Since 2011, he has been a Full Professor with the Department of Energy Technology, Aalborg University, Denmark, where he is responsible for the Microgrid Research Program. Since 2014, he has been the Chair Professor of Shandong University; since 2015, he has been a Distinguished Guest Professor with Hunan University; since 2016, he has been a Visiting Professor Fellow at Aston University, U.K.; and a Guest Professor at the Nanjing University of Posts and Telecommunications. Since 2019, he has been a Villum Investigator with Villum Fonden, which supports the Center for Research on Microgrids (CROM), Aalborg University, where he has been the Founder and the Director. He has published more than 600 journal articles in the fields of microgrids and renewable energy systems, which are cited more than 60 000 times. His research interests include different microgrid aspects, including power electronics, distributed energy-storage systems, hierarchical and cooperative control, energy management systems, smart metering, and the Internet of Things for AC/DC microgrid clusters and islanded minigrids. Specially focused on microgrid technologies applied to offshore wind, maritime microgrids for electrical ships, vessels, ferries and seaports, and space microgrids applied to nanosatellites and spacecrafts. In 2015, he was elevated as IEEE Fellow for his contributions on distributed power systems and microgrids. He received the Best Paper Award of the IEEE TRANSACTIONS ON ENERGY CONVERSION from 2014 to 2015 and the Best Paper Prize of the IEEE-PES, in 2015. He received the Best Paper Award of the *Journal of Power Electronics*, in 2016. During seven consecutive years, from 2014 to 2020, he was awarded by Clarivate Analytics (former Thomson Reuters) as a highly cited researcher with 50 highly cited articles. He is an associate editor for a number of IEEE Transactions.