*Proceedings of the International Conference on Industrial Engineering and Operations Management*

Publisher: IEOM Society International, USA DOI: 10.46254/AN15.20250203
Published: February 18, 2025

# Securing Mobile Connectivity: A Data-Driven Approach for Detection of Electricity Theft and Anomalies in 4G and 5G Telecommunications Base Stations

**Malebo L.G Maboko**
M.Sc (Eng) Student
School of Mechanical, Industrial, and Aeronautical Engineering
University of the Witwatersrand (Wits)
Johannesburg, South Africa
672871@students.wits.ac.za

**Mncedisi Dewa, Ph.D.**
Senior Lecturer
School of Mechanical, Industrial and Aeronautical Engineering
University of the Witwatersrand (Wits)
Johannesburg, South Africa
mncedisi.dewa@.wits.ac.za

## Abstract

This study presents a framework centered on machine learning and deep neural networks (DNN) designed to develop a robust system to detect power theft and electrical anomalies in next-generation telecommunication base stations. Using a labelled dataset of non-fraudulent power consumption records with 11 features, this research employed supervised machine learning algorithms, including logistic regression, random forest, and DNN. Comprehensive data preprocessing, feature engineering, model training, evaluation, and optimization were carried out to ensure effective model performance. The results demonstrated the superiority of the Random Forest model, achieving perfect accuracy (100%) and zero root mean squared error (RMSE) across datasets with outliers injected at 25%, 50% and 75%. The DNN model achieved comparable performance on cleaner datasets (100% accuracy and 0% RMSE on the 25% outlier dataset) but showed reduced robustness as noise levels increased, with accuracy dropping to 83.32% and 75.11% on the 50% and 75% outlier datasets, respectively. Logistic regression, while simpler and interpretable, struggled to handle datasets with high levels of outliers, achieving 74.92% accuracy on the 75% outlier dataset with an RMSE of 0.5008. The paper also provides information on the potential of hybrid models and advanced preprocessing techniques to enhance future anomaly detection systems.

## Keywords

Classifier, DNN, Ensemble learning, Hyper-parameter tuning, Non-Technical losses

## 1. Introduction

The power grid experiences losses broadly categorised into technical and non-technical losses. Technical losses occur due to inherent power dissipation in the physical components of the power system, such as internal electrical resistance in generators and transformers, as well as through transmission lines and other connecting elements. These losses are a natural consequence of electrical resistance and inefficiencies within the system infrastructure (Smith, 2024).

Non-technical losses (NTLs) refer to losses that are external to the power distribution system and are not caused by the physical characteristics of the network. NTLs typically arise from issues such as theft of electricity, meter tampering, billing inaccuracies, and unauthorized connections (Johnson, 2024). These losses represent a significant challenge for power utilities, as they impact revenue and operational efficiency (Brown, 2024).

NTLs typically arise from issues such as illegal connections, fraudulent acts, and billing irregularities. Illegal connections involve unauthorized access to the power grid, leading to unmetered and unaccounted electricity usage (Johnson, 2024). Fraudulent acts include tampering with meters to reduce recorded consumption or bypassing meters entirely, which directly impacts the accuracy of billing (Brown, 2024). Billing irregularities can occur due to administrative errors, faulty meters, or deliberate under-reporting of usage, all contributing to revenue losses for utilities (Davis, 2024).

The methods used to detect NTLs can be categorized as follows:

- Theoretical investigations;
- Hardware-based approaches;
- Software approaches;
- Integrated hardware and software solutions and
- Manual inspections for tampering and component integrity checks.

In this paper, these methods are studied and extended to include the detection of electrical theft and electrical equipment malfunctions within the telecommunications industry.

The remainder of the paper is organized according to the flow chart shown in Figure 1 and is detailed as follows: Section 2 presents the literature review. Section 3 details data collection and preprocessing. Fraudulent users that were modeled and injected into the data set as well as feature engineering are discussed in Section 4. Model building and optimization are covered in Section 5. The performance of the classifier is presented in Section 6. Section 7 presents future recommendations, and finally, Section 8 details the conclusions.

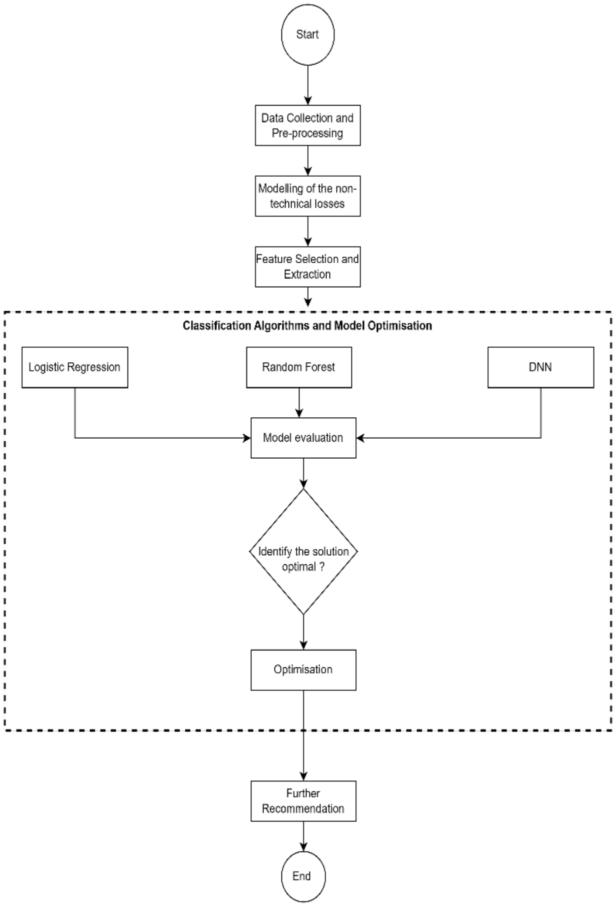
Start
Data Collection and Pre-processing
Modelling of the non-technical losses
Feature Selection and Extraction
Classification Algorithms and Model Optimisation
Logistic Regression
Random Forest
DNN
Model evaluation
Identify the solution optimal ?
Optimisation
Further Recommendation
End

Figure 1. Overview of the solution.

## 1.1. Objectives

The primary aim of this study is to prototype a highly effective data-driven machine learning-based approach for early detection of power theft and electrical anomalies in new generation telecommunications base stations.

- Investigation and evaluation of existing electricity theft and electrical anomaly detection schemes.
- Design and implementation of a new electricity theft technique in new-generation telecommunications base stations.

## 1.2 Problem statement

Electrical anomalies in base stations and theft of electricity in the power grid negatively affect the quality of service in mobile networks. This consequently leads to the loss of customers in telecommunication companies and overall poor mobile communication in the community.

## 1.3 Requirements

- The solution should be capable of integrating diverse data sources, such as energy consumption, via an application programming interface (API). Primarily, the solution should be able to ingest CSV data from HUAWEI's NetECO power monitoring platform.
- The prototype should also incorporate preprocessing techniques to clean and normalise the data, ensuring its suitability for exploratory data analysis.
- Effective feature engineering is vital for the implementation of accurate machine learning models. The system should include feature extraction methodologies, as well as techniques for selecting the most informative features for model performance optimisation.

- The prototype should involve training and testing machine learning models using labeled data to detect patterns indicative of suspected power theft or electrical anomalies. Furthermore, the solution should be ready for deployment in near-real-time environments to enable continuous monitoring and detection of suspicious activities while maintaining scalability and efficiency in operational settings.

### 1.4 Assumptions and Constraints

The solution assumes the availability of sufficient data that could be labeled to train machine learning models. These data should be accurately modeled to represent instances of suspicious and loyal electricity consumption, including cases of electrical equipment malfunctions. The assumptions include consistent data quality, adequate data storage capacity, computing memory, and processing power, as well as reliable data transmission mechanisms. However, constraints such as data inconsistencies, network failures, and limited data storage capacity may affect the performance of the solution. The solution operates following legal and regulatory frameworks that govern data privacy, consumer rights, and law enforcement protocols. Assumptions encompass strict adherence to applicable regulations and compliance with industry norms regarding data management and security practices. However, constraints such as legal intricacies, evolving regulatory landscapes, and jurisdictional disparities can present challenges to the execution and functionality of the solution.

### 1.5 Success criteria

The solution should demonstrate an accuracy of at least 95% in identifying instances of power theft and potential malfunctioning equipment, achieving a low RMSE to ensure reliable detection. The solution should provide timely detection of suspicious activities, enabling swift intervention and mitigation measures to prevent network interruptions. The solution should be scalable to handle increasing volumes of data and growing network infrastructure, ensuring its effectiveness and applicability as telecommunication networks expand over time.

## 2. Review of the literature

According to Smith (2004). Electricity theft within the power distribution network presents a significant challenge, often manifesting itself through two primary methods: tampering with energy meters and other electrical components in the grid and unauthorized connection of energy lines. Such illicit activities not only lead to revenue losses for utility providers, but also pose safety risks to both consumers and infrastructure. This paper aims to comprehensively address the issue of theft of electricity in the new generation base station and malfunctioning electrical equipment in both of these areas. The adverse effects of electricity theft on the telecommunication infrastructure are profound, both in South Africa and worldwide. In South Africa, the telecommunications sector bears the brunt of theft of electricity through various channels, including illegal connections and vandalism. This leads to disruptions in telecommunications services, hinders communication networks, and causes financial losses for service providers (Ntuli et al., 2018).

Furthermore, the proliferation of electricity theft undermines efforts to expand and upgrade the telecommunication infrastructure in developing countries. Instead, the scarce resources that could be allocated to infrastructure development are diverted towards combating electricity theft and repairing damage caused by illegal activities (Farhangi, 2010).In South Africa, where electricity theft is rampant, utility providers have been actively combating this issue. Cabinet intrusion, a form of electricity theft involving unauthorised access to electrical cabinets or distribution boxes, is prohibited by regulations set forth by regulatory bodies such as the National Energy Regulator of South Africa (NERSA). According to Eskom Holdings SOC Ltd (2022), electricity theft is illegal in South Africa. This crime leads to the following.

- Extraction of energy meters and other crucial electrical apparatus;
- Unauthorized connections;
- Damage to infrastructure;
- Manipulation of energy meter readings.

The illegal removal of energy meters and other electrical components constitutes a global issue. This type of theft of electrical equipment disrupts the power supply within communities and results in voltage fluctuations, causing interruptions in household power services. Furthermore, unlawful removal of vital power equipment, such as backup batteries, compromises the resilience of the energy network infrastructure. For example, during removal of the load reduction schedule, there is an immediate power loss, exacerbating service reliability concerns (Glauner et al., 2017). Illegal connections represent another facet of electricity theft, leading to unaccounted-for power usage by electricity distribution companies. Vandalism, including the destruction of electrical equipment, can alter meter readings, introduce billing discrepancies, and disrupt energy meter communication. Additionally, tampering with energy meters, such as bypassing them, constitutes fraudulent behaviour that results in electricity theft. This manipulation enables customers to consume electricity while registering zero-meter readings, thus avoiding accurate billing.

## 2.1. Related Work

Electricity theft and anomaly detection are critical in the power sector. Logistic regression (LR) is simple and interpretable, but limited to nonlinear data (Akhavan-Hejazi & Mohsenian-Rad, 2018). K-Nearest Neighbors (KNN) is effective for small datasets but computationally intensive for large ones (Jindal, Singh, & Agarwal, 2016). Support Vector Machines (SVM) handle high-dimensional spaces well, but require careful parameter tuning (Akhavan-Hejazi & Mohsenian-Rad, 2018). Random Forest (RF) is robust to overfitting and handles large datasets efficiently (Jindal, Singh, & Agarwal, 2016). Deep neural networks (DNNs) excel in learning complex patterns but need substantial resources (Zheng et al., 2018).

The study focuses primarily on software-based solutions. These non-hardware solutions can be divided into three main categories: game theory, estimation, and classification. The majority of solutions within this category utilize classification techniques, including rule-based systems (RBS), XGBoost, support vector machines (SVM), logistic regression, optimal path forest, k nearest neighbors (kNN), decision trees (DT) and artificial neural networks (ANN). Section 5 will present the classification algorithms chosen for the theft detection scheme in this work.

# 3. Data collection and pre-processing

The raw data set was obtained from NetECO HUAWAI energy monitoring software for telecommunication base stations, using Rain Networks Company private data for a triple-stacked base station. This database consists of energy consumption data collected from January 2016 to December 2016. A triple stack telecommunications base station has a maximum power consumption of 6.5kW at 100% load of UE usage and 4100kW at 0% load of UE usage. A triple-stack radio configuration 3 sector (5G-3600MHz) + 3 sector 4G (700MHz+2600MHz). The dataset comprises non-fraudulent power consumption records with 11 features. Further data filtering will be done to achieve a complete and non-fraudulent dataset. This will be done by removing outliers and customers with missing values and correcting customers with inconsistent timestamps.  Figure 2 shows the average of total energy consumption. Figure 3 below presents the distribution of the non-fraudulent data. The data set contains 503879 rows of energy usage in kWh.

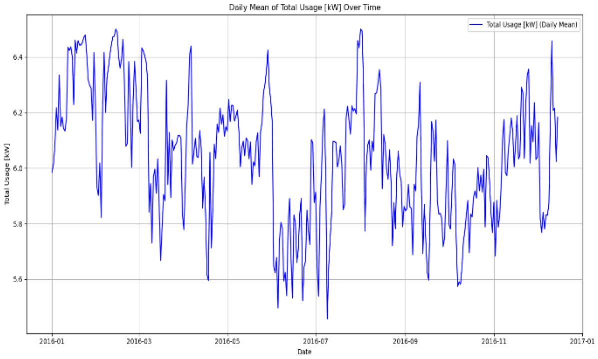

Figure 2. Daily average of total energy usage

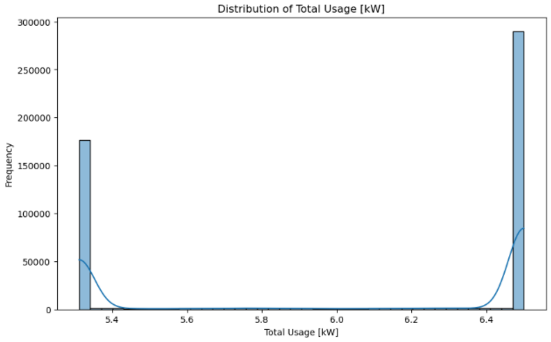

Figure 3. Distribution of nonfraudulent data

# 4. Modeling of non-technical losses dataset and feature engineering

To train and validate the NTL detection algorithm, it is essential to include fraudulent samples in the data set. The data set used, sourced from rain networks, comprises data collected from a triple-stack telecommunication base station that had no alarms. Consequently, it is assumed that this data set does not contain any inherent fraudulent behaviour. To create a more representative dataset, fraudulent samples were artificially generated using two methods: the physical attack technique and the data attack technique (S. Mclaughlin et al.). Four different types of fraudulent profiles were considered for modelling the NTLs, all modifications were based on these. The modelled profiles are described as follows:

- Zero-consumption reading by bypassing the energy meter to the base station (data attack/ physical attack technique).
- Zero consumption reading by tempering with an energy meter to the base station (physical attack technique).
- The total value of the stolen components/wires is different from the maximum energy consumption or less than the minimum consumption power (physical attack technique).
- Hooking of the main meter to the power station, and power line to residences, thereby increasing the average consumption (physical attack technique).

Feature engineering enables the management of high-dimensional time-series data by utilizing characteristics that describe the dataset rather than relying on the raw data itself. This process simplifies the data, making it easier to analyze and interpret by reducing its dimensionality while retaining essential information. By extracting relevant features, one can improve the efficiency and accuracy of classification tasks. Feature extraction helps mitigate issues related to noise and redundancy in the data, thus enhancing the overall quality of the classification models (Brown, 2023). Of the 11 features, the total power consumption was chosen to represent the data using the PCA (principle component analysis). The outlier parameter is determined using the z-score technique. Identifies outliers by measuring how many standard deviations a data point is from the mean of the data set. The steps involved in calculating the z score and determining outliers are presented below. It is based on quantile analysis, as detailed in (John et al.).

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1}$$

Where $\mu$ is the mean, $x_i$ is the data points and N is the total number of data points.

$$\sigma = \sqrt{\frac{1}{N}\sum_{I=1}^{N}(x_i - \mu)^2} \tag{2}$$

Where standard deviation ($\sigma$) represents a measure of the dispersion of the data points.

$$z_i = \frac{x_i - \mu}{N} \tag{3}$$

Where Z-Score ($z_i$) represents the number of standard deviations that a data point is from the mean. For this task $|z_i| > 3$ as recommended by (John, G et al.).

Further feature generation was performed to produce a resultant target column for machine learning and neural network classification. Subsequently, the profiles were engineered to resemble fraudulent consumption patterns. Three main types of consumers were modelled. Fraudulent activities were named outliers. The first data set was injected with 25% fraudulent activities, the second data set was injected with 50% fraudulent activities, and the third data set was injected with 75% fraudulent activities. Figure 4 shows the imbalance ratio in terms of binary targets.

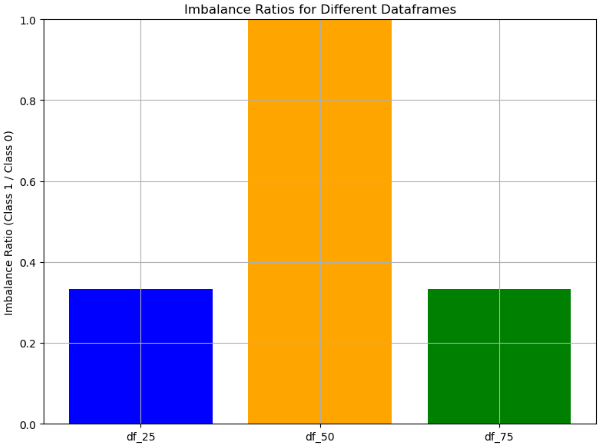

Figure 4. Imbalance ratio for the different datasets

$N_0$ and $N_1$ represent the number of samples in the minority and majority classes, respectively, in a binary classification problem. The imbalance ratio $R$ can be expressed as:

$$R = \frac{N_{minority}}{N_{majority}} \tag{4}$$

## 5 Machine learning, DNN implementation, and optimization

All source code used for this project can be found on this platform:
https://drive.google.com/file/d/1Ybga_hoc9tYAdCT4uxysFYXjUgny6nFk/view?usp=sharing

In the realm of fraud detection, selecting the appropriate machine learning models is crucial to developing a robust and effective system. Logistic regression, random forest, and deep neural networks (DNN) are the algorithms used in this study and are the most effective techniques used due to their unique advantages and capabilities.

Logistic regression is often chosen for its interpretability and efficiency. It provides a clear understanding of how each input feature affects the output, making it invaluable for identifying patterns in fraudulent behavior. As a binary classification tool, it serves as a strong baseline model, efficiently handling large datasets, and performing well in real-time fraud detection scenarios (Hastie, Tibshirani, & Friedman, 2009). This methodology was used in this report for hyper parameters. The key hyper-parameters set are as follows:

- solver=lbfgs': This specifies the optimization algorithm to use, which in this case is the limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm.
- max_iter=1000: This sets the maximum number of iterations for the optimization algorithm.
- random_state=42: This ensures reproducibility by setting a seed for random number generation.

By setting these hyperparameters, the logistic regression model can effectively learn from the data and make accurate predictions while also preventing overfitting through regularization.

$$\mathrm{P}(\mathrm{y} = 1 \mid \mathrm{X}) = \sigma(\mathrm{X} \cdot \beta) \tag{5}$$

Where $\mathrm{P}(\mathrm{y} = 1 \mid X)$ is the probability that the output $y$ is 1 given the input $X$, σ is the sigmoid function $\sigma(\mathrm{z}) = \frac{1}{1+e^{-z}}$. $X$ is the input feature vector. β is the vector of coefficients (weights) for the input features.

In logistic regression, hyperparameters, such as the regularization term, play a crucial role. Regularization helps to prevent overfitting by penalizing large coefficients. The regularization term is typically added to the cost function J(β) L2 regularization.

$$\mathrm{J}(\beta) = -\frac{1}{m}\sum_{1}^{m}[y_i \log\sigma(X_i \cdot \beta) + (1 - y_i)\log(1 - \sigma(X_i \cdot \beta))] + \frac{\lambda}{2m}\sum_{j=1}^{n}\beta_j^2 \tag{6}$$

where $m$ is the number of training examples. $y_i$ is the actual label for the i-th training example. $X_i$ is the feature vector for the i-th training data point. λ is the regularization parameter. $n$ is the number of features.

The goal is to minimize the cost function J(β). This is typically done using gradient descent, where the update rule for the coefficients β is:

$$\beta := \beta - \alpha\nabla\mathrm{J}(\beta) \tag{7}$$

Where α is the learning rate and ∇J(β) is the cost function gradient with respect to β. The cost function gradient is as follows:

$$\nabla J(\beta) = \frac{1}{m}X^T(\sigma(X \cdot \beta) - y) + \frac{\lambda}{m}\beta \tag{8}$$

This equation iteratively adjusts the coefficients to minimize the $J(\beta)$

Random Forest stands out due to its robustness and ability to handle categorical and continuous data. This ensemble learning method combines multiple decision trees to enhance accuracy and reduce overfitting (Breiman, 2001).

Additionally, random forest models provide insight into the importance of features, helping to understand the factors that contribute to fraud. Their versatility and ability to capture nonlinear relationships make them highly effective in various fraud detection tasks.

Deep neural networks (DNN) excel at recognizing complex patterns and relationships within data. They are particularly suited for detecting sophisticated fraudulent activities that may not be easily captured by simpler models. DNNs can be scaled and adapted to improve performance as the dataset grows in size and complexity (Goodfellow, Bengio, & Courville, 2016).

The processed data will be split into the validation train and test data. The 80-20 split was used, meaning that 80% of the data is used for training and 20% for testing. During cross-validation, the training data are divided into training and validation sets. In 5-fold cross-validation, the training data is split into 5 parts. Each part is used as a validation set once, while the remaining 4 parts are used for training. Hyperparameter tuning optimization was applied to improve the performance of logistic regression. The Hyper parameter tuning optimization methods were not applied for the random forest as it is an ensemble learning algorithm.

## 6. Results and Discussion

This study evaluated the performance of logistic regression, random forest, and deep neural network (DNN) in data sets that contained varying proportions of outliers injected (25%, 50% and 75%). Performance was assessed using precision, root mean squared error (RMSE) and confusion matrices. In the following, we provide a detailed numerical analysis of the models' performance and additional insights. The accuracy and RMSE are depicted in Figures 5 and 6 respectively, while the numerical results of the accuracy, rmse, and the confusion matrix are shown in Table 1.

### 6.1. Accuracy

Random Forest consistently achieved perfect accuracy across all datasets, with accuracy rates of 100% for the 25%, 50%, and 75% outlier proportions. This highlights Random Forest's ability to maintain robustness and resilience to varying levels of noise due to its ensemble-based approach.

DNN performed strongly in the dataset with 25% outliers, matching Random Forest's perfect accuracy of 100%. However, its performance declined with higher levels of outliers, with accuracy dropping to 83.32% for the 50% data set and 75.11% for the 75% data set. This decline indicates that while DNN can generalise well, it is more sensitive to noise in data compared to Random Forest.

Logistic regression exhibited the lowest precision, steadily decreasing from 83.29% for the 25% dataset to 79.14% for the 50% dataset, and further to 74.92% for the 75% dataset. Its simpler linear nature limits its adaptability to the complex patterns introduced by outliers, making it less effective under noisy conditions.

### 6.2. RMSE

Random Forest maintained an RMSE of 0.0000 across all datasets, which means perfect predictions and no error distribution regardless of the outlier levels. This further solidifies its reliability in handling noisy datasets.

DNN showed low RMSE in the data set with 25% outliers (0.0000), but exhibited higher values as noise increased. RMSE increased to 0.4084 for the 50% dataset and 0.4986 for the 75% dataset, reflecting its increasing difficulty in minimizing prediction errors in noisy conditions.

Logistic regression displayed the highest RMSE in all datasets, starting at 0.4088 for the 25% dataset, increasing to 0.4567 for the 50% dataset, and peaking at 0.5008 for the 75% data set. These results emphasize its reduced ability to handle data sets with significant proportions of outliers effectively.

### 6.3. Confusion matrix

The confusion matrices provide additional insight into model performance by breaking down the predictions into true positives, true negatives, false positives, and false negatives.

Random Forest consistently classified all samples correctly across all datasets, with no false positives or false negatives observed. For example, in the 50% outlier dataset, Random Forest achieved 100,602 true positives and

100,950 true negatives, which did not result in misclassifications. This demonstrates its unparalleled precision and robustness even in highly noisy data sets.

DNN revealed some misclassifications as the outlier proportion increased. For example, in the 75% outlier dataset, DNN correctly classified 50,551 true positives and 100,885 true negatives. However, it also misclassified 50,116 samples as false positives, indicating a struggle to maintain precision in datasets with high noise levels.

Logistic regression exhibited the poorest performance in terms of misclassifications, particularly in the 75% outlier dataset, where it failed to correctly classify any true positives or true negatives. This highlights its inability to cope with substantial noise, further supported by its higher RMSE values (Table 1).

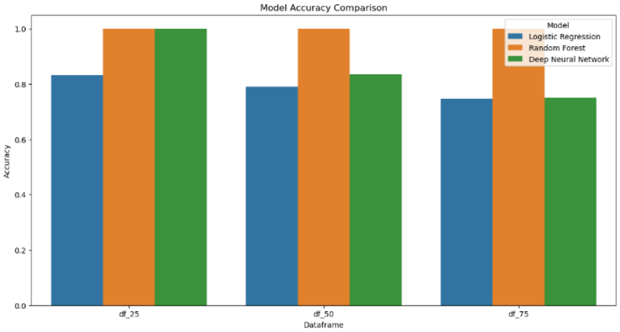

Figure 5. Shows the accuracy results

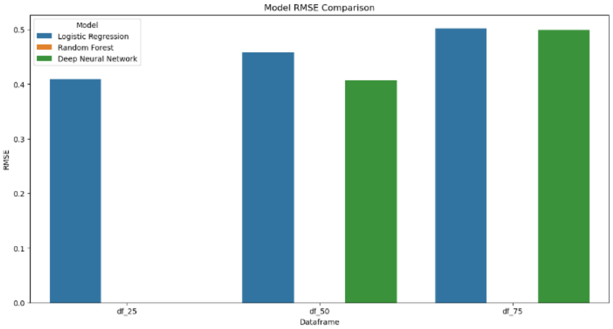

Figure 6. Shows RMSE results.

Table 1. Summary of Model Evaluations

| **Model** | **Dataframe** | **Mean Accuracy** | **Std Dev** | **Test Accuracy** | **Confusion Matrix** | **RMSE** |
|---|---|---|---|---|---|---|
| Logistic Regression | df_25 | 0.8333 | 0.0005 | 0.8329 | 151083 0<br>33687 16782 | 0.408825 |
| | df_50 | 0.7918 | 0.0004 | 0.7914 | 92177 8425<br>33613 67337 | 0.456696 |
| | df_75 | 0.7502 | 0.0000 | 0.7492 | 0 50551<br>0 151001 | 0.500808 |
| Random Forest | df_25 | 1.0000 | 0.0000 | 1.0000 | 151083 0<br>0 50469 | 0.000000 |
| | df_50 | 1.0000 | 0.0000 | 1.0000 | 100602 0<br>0 100950 | 0.000000 |
| | df_75 | 1.0000 | 0.0000 | 1.0000 | 50551 0<br>0 151001 | 0.000000 |
| Deep Neural Network | df_25 | 1.0000 | 0.0000 | 1.0000 | 151083 0<br>0 50469 | 0.000000 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | df_50 | 0.8333 | 0.0005 | 0.8332 | 100602 0<br>33613 67337 | 0.408376 |
| | df_75 | 0.7513 | 0.0007 | 0.7511 | 50551 0<br>50116 100885 | 0.498649 |

## 7. Future recommendations

For future work, several recommendations can enhance the performance and applicability of the models evaluated. Hyperparameter tuning should be prioritized, including optimizing parameters like tree depth and the number of estimators for Random Forest, as well as learning rates, batch sizes, and network architectures for Deep Neural Networks (DNNs). Incorporating explainability tools such as SHAP (SHapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations) can help identify the most influential features and build trust in model predictions. Clustering techniques such as k-means or DBSCAN can be employed to preprocess datasets, identify patterns in outlier distributions, or detect noise, further improving data quality before training. Furthermore, the development of hybrid models, such as random forest stacking and DNN, could balance their respective strengths, achieving robust performance under diverse conditions. Practical implementation should consider leveraging Random Forest for high-stakes tasks requiring robustness while utilizing DNNs for applications that demand generalization on complex patterns. Finally, integrating AutoML pipelines for automated hyperparameter tuning and exploring advanced outlier detection methods like Isolation Forest or Elliptic Envelope could significantly improve model performance and reliability.

## 8. Conclusions

This paper delves into comparative analysis of logistic regression, random forest, and DNN on datasets with varying levels of outliers provided significant insight. This study aims to prototype a highly effective data-driven DNN-based approach for early detection of power theft and electrical anomalies in the new generation telecommunications base stations. Random Forest emerged as the most reliable model in terms of accuracy across all outlier levels, while DNN showed competitive performance with lower RMSE in noisier datasets than logistic regression. Logistic regression, being a simpler model, struggled with the complexity introduced by outliers. The study highlights the importance of the selection of electrical theft and anomaly detection models based on the characteristics of the dataset, and underscores the potential benefits of hybrid models and advanced preprocessing techniques. Future work can focus on optimizing these approaches and exploring new methodologies to further enhance model robustness and accuracy.

## References

Akhavan-Hejazi, H., & Mohsenian-Rad, H., Power systems big data analytics: An assessment of paradigm shift barriers and prospects. *Energy Reports*, 4, 91-100, 2018.

Breiman, L., *Random Forests*. *Machine Learning*, 45(1), 5-32, 2001.

Brown, L., 'Understanding Non-Technical Losses in Power Grids'. 2024. Available at: https://www.researchgate.net/publication/Understanding_Non-Technical_Losses_in_Power_Grids (Accessed: 22 July 2024).

Davis, K., 'Mitigation of Non-Technical Losses in the Power Sector'. 2024. Available at: https://www.ieee.org/publications/magazines/power-and-energy/mitigation-of-non-technical-losses-in-the-power-sector (Accessed: 22 July 2024).

Goodfellow, I., Bengio, Y., & Courville, A., *Deep Learning*. MIT Press. 2016.

Gupta, S., Kumar, A., & Singh, A., Addressing the Challenge of Electricity Theft: A Comprehensive Review. International Journal of Electrical Power & Energy Systems, 105, 347-360, 2019.

Hastie, T., Tibshirani, R., & Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer., 2009.

https://www.eskom.co.za/distribution/customer-service/public-safety/, *Electricity Safety Tips*, Accessed 29/09/2022

Jindal, A., Singh, N., & Agarwal, K., Analytics framework for detecting electricity theft. *Computing Research Repository (CoRR)*, abs/1611.06609, 2016.

John, G., and Smith, L., "Quantile Analysis for Outlier Detection." *Journal of Statistical Methods*, 15(3), 234-250, 2022.

Johnson, R., 'Power Losses in Electric Grids'. 2024. Available at: https://www.sciencedirect.com/topics/engineering/power-losses (Accessed: 22 July 2024).

Li, C., Ma, X., Li, H., & Zhang, Y., Research on Detection of Illegal Electricity Usage Based on Energy Consumption Characteristics. Journal of Physics: Conference Series, 1529(2), 022087, 2020.

National Energy Regulator of South Africa (NERSA). (n.d.). Regulations on Unauthorized Use of Electricity. Retrieved from https://www.nersa.org.za/,Accessed 02/18/2014.

P. Glauner, J. A. Meira, P. Valtchev, R. State, and F. Bettinger, "The challenge of non-technical loss detection using artificial intelligence: A survey," *International Journal of Computational Intelligence Systems*, vol. 10, no. 1, p. 760–775, 2017.

Smith, J., 'Technical Losses in Electrical Power Systems'. 2024, Available at: https://www.electrical4u.com/technical-losses-in-electrical-power-systems/ (Accessed: 22 July 2024).

Smith, T.B., Electricity Theft: A Comparative Analysis. Energy Policy, 32, 2067-2076, 2004.

Wiedmann, T., & Rooker, T., Electricity Theft: A Comparative Analysis of Prepaid and Postpaid Metering in South Africa. Energy Policy, 129, 1175-1183, 2019.

Zheng, Z., Yang, L., Niu, X., Dai, H., & Zhang, W., Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids. *IEEE Transactions on Industrial Informatics*, 14(4), 1606-1615, 2018.

## Biographies

**Malebo Maboko** is currently a Senior Data Scientist at CSIR. He is an MSc in Engineering, specialising in Artificial Intelligence student at the School of Mechanical, Industrial and Aeronautical Engineering, University of the Witwatersrand. He currently holds a B.Sc in Engineering in Electrical Engineering: (Information Engineering) from the School of Electrical and information Engineering, University of the Witwatersrand. He previously served as a Special Project Research Engineer at rain Networks, South Africa. Candidate Control, Instrumentation and Automation Engineer at BBE and Proconics Groups. His research interests are in Telecommunications Engineering, IoT, Data Analytics and Machine Learning. He is currently a registered as a Candidate Engineer with the Engineering Council of South Africa (ECSA).

**Mncedisi Dewa** holds a PhD in Industrial Engineering from Stellenbosch University, a Master of Engineering degree in Manufacturing Systems and Operations Management and BEng degree in Industrial and Manufacturing Engineering from the National University of Science and Technology, Zimbabwe. He is currently a Senior lecturer in the School of Mechanical, Industrial and Aeronautical Engineering at the University of the Witwatersrand.
Within the School, he conducts teaching, learning, postgraduate supervision, academic consultation and collaborative research. He has served as an academic for over twelve years now, with seven of those years in the Higher and Tertiary education sector in South Africa. His research interests are in the areas of E-learning, Digital Assistance Systems, system modeling, simulation and Engineering Education. He is a registered member of the South African Institute of Industrial Engineers (MSAIIE), the South African Society for Engineering Education (MSASEE) and a registered Candidate Engineer with the Engineering Council of South Africa (ECSA).