ARCHIVES OF ELECTRICAL ENGINEERING VOL. 73(4), pp. 1103–1121 (2024)

DOI 10.24425/aee.2024.152113

Short-term prediction of power outages in electrical distribution networks

PFANELO MAKUNGO, TLOTLOLLO HLALELE ✉, MBUYU SUMBWANYAMBE

University of South Africa, Department of Electrical and Smart Systems Engineering
Johannesburg, South Africa
e-mail: 42494680@mylife.unisa.ac.za, {✉ hlalets/sumbwm}@unisa.ac.za

(Received: 25.03.2024, revised: 24.11.2024)

Abstract: Predictive maintenance and reliability engineering are critical in industrial settings to enhance equipment performance and minimize unplanned downtime. This research, conducted within the machine learning framework, presents innovative solutions to the challenging problem of equipment failure prediction. The study creatively utilizes extensive datasets, including equipment records, weather conditions, and maintenance logs, to develop robust predictive models. Two distinct machine learning models are established for equipment and cables/lines, addressing the intricacies of class imbalances and missing data attributes. Model refinement, feature engineering, and interdisciplinary collaboration enhance predictive accuracy, precision, and recall. Notably, this research highlights the creative application of engineering knowledge and data science techniques, reasoning about complex equipment systems, and the importance of problem decomposition. The outcomes underscore the potential for real-time predictive maintenance in industrial contexts, offering substantial cost savings and improved equipment reliability. This research contributes to the evolving field of predictive maintenance and paves the way for future innovations in reliability engineering.

Key words: data preprocessing, load-shedding, machine learning, outage prediction, predictive maintenance, predictive model

1. Introduction

Medium voltage networks are a critical component of power grids, providing electricity to homes, businesses, and industries. However, these networks are vulnerable to various factors that can cause power outages, including extreme weather events, equipment failures, and overloading. Asaridis *et al.* (2021) highlight the susceptibility of these systems to flood-related disruptions, emphasizing the need for robust predictive models [1]. These outages can have significant economic and social impacts, leading to lost productivity, disrupted services, and even safety hazards. Nduhuura *et al*. (2021) discuss the profound effects of such outages on urban households in developing countries, stressing the urgency of addressing these challenges [2].

Exploring new technologies and methods to prevent power outages in medium-voltage networks is essential to address this. Onaolapo *et al*. (2022) provide a comparative assessment of traditional and AI-based forecasting methods, showing how artificial intelligence can significantly enhance the reliability and efficiency of these networks [3]. This improvement is critical not only for reducing the risk of outages but also for enhancing customer service.

Despite ongoing efforts to maintain and improve the reliability of the medium voltage distribution network, power outages continue to occur. O'Fallon and Gopstein (2021) quantify the resilience benefits that smart grid technologies offer, yet they acknowledge that outages remain a significant problem [4]. Thus, there is a clear need to develop a more proactive approach to outage prevention, utilizing the vast amount of data available to utilities to predict and prevent potential outages, as discussed by Kiran D. (2017) in their analysis of failure modes and effects [5].

By using machine learning algorithms to analyze weather patterns, equipment maintenance logs, and other relevant factors, it is possible to accurately predict potential power outages in distribution networks and prevent them before they occur. Tsioumpri *et al*. (2021) demonstrate the efficacy of weather-related fault prediction in minimally monitored distribution networks, showcasing the predictive power of these algorithms [6].

For the City of Cape Town Utility, developing a machine learning algorithm to predict and prevent power outages in distribution networks can lead to significant cost savings by reducing the frequency and duration of outages, as noted by Fakih *et al*. (2020) who explore the economic impacts of power reliability on manufacturing firms [7]. Additionally, such advancements can improve the reliability and safety of the power grid and reduce the need for emergency repairs and maintenance, thereby helping to minimize the industry's environmental impact while maintaining a reliable power supply, as Luo *et al*. (2023) suggest in their study on intelligent decision-making methods for outage maintenance [8].

The research aims to achieve several key objectives. First, it seeks to gather and analyze pertinent data concerning weather patterns, equipment maintenance, and other factors that have the potential to trigger power outages within the network. Second, the study intends to identify the most pertinent and highly correlated features associated with power outages, subsequently converting them into a format suitable for machine learning algorithms. Third, the research involves the development and training of a machine learning algorithm, utilizing the selected features and historical data on power outages in the network. Finally, the study will assess the performance of the machine learning algorithm using relevant metrics and make any necessary refinements to enhance its effectiveness.

2. Problem formulation

A supervised learning problem was addressed through the formulation of two distinct machine learning models. The research focused on predicting equipment and cable/line failures using labelled datasets, applying a supervised learning approach. The formulation process involved analyzing data from 1521 recorded incidents of equipment failures occurring between 1 January 2018 and 30 May 2023. The data for this research was collected from the eastern area, one of the regions within the City of Cape Town municipality in South Africa. This specific geographic area served as the primary source of incident data related to the outages. The heat map in Fig. 1 visualizes hot spots within the dataset. These hot spots represent areas with a higher concentration of outages, providing valuable insights into the spatial distribution of incidents within the eastern region of the City of Cape Town municipality.

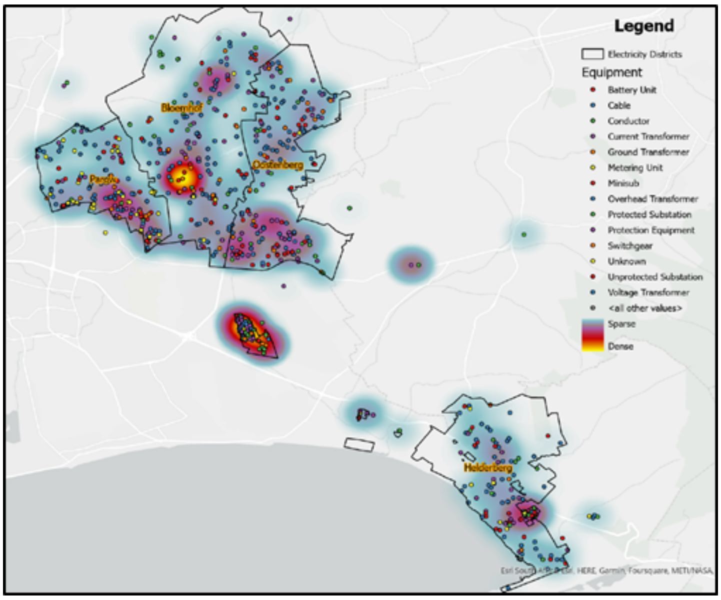

Fig. 1. Geocoded outages arranged per district

The problem of recognizing potential equipment failures based on environmental parameters is framed as a classification task. Two separate models are developed to tackle the classification task. The first model, aimed at predicting equipment failures, involved a multi-class classification problem. The classes utilized are indicated in Table 1. The model needed to differentiate between these equipment categories to make precise predictions. The second model also involved multi-class classification, predicting cable and line issues. By categorizing these problems into distinct classes, as shown in Table 1, the model could determine the specific issue affecting the cable or line, facilitating targeted maintenance and repairs. The decision to create two distinct models, one for equipment failure prediction and the other for cable/line issue prediction, was driven by data attributes and the challenges posed by missing information. Specifically, the lack of key attributes, such as age and manufacturer details for cables and overhead lines, made it challenging to predict failures accurately for these equipment types. Thus, a decision was made to develop separate models to leverage the available data efficiently without the need for predictive imputation or potentially inaccurate estimations.

Table 1. Class definitions of equipment

Class description	**Class #**	**% Count**
Cable	2	32.38%
Overhead conductor	3	18.43%
Ground transformer	5	0.97%
Minisub	8	6.29%
Overhead transformer	9	1.20%
Protected substation	10	2.17%
Protection equipment	11	22.17%
Switchgear	12	2.92%
Unknown	13	6.29%
Unprotected substation	14	1.65%

3. Dataset

The research's strength lies in integrating historical equipment data, asset management records, historical outage data, and external weather data. These datasets collectively formed the basis for building machine-learning models that could predict equipment failures and analyze outage patterns. By combining spatial, temporal, equipment, and environmental factors, the research aimed to provide valuable insights into the complex problem of outage prediction and asset management in the given area.

3.1. Historical recorded outage data

The historical outage data used for this analysis was sourced from outage report SAP database records maintained by the City of Cape Town Utility, encompassing five years from January 2018 to May 2023. The data cleaning and preprocessing phase was initiated to ensure the data's quality and consistency, which involved data validation to identify and correct inconsistencies or outliers and standardization of timestamps for temporal analysis. Outage locations were mapped geospatially using geographic information system (GIS) tools, providing a spatial context to each equipment item. This mapping facilitated spatial analysis by associating outages with specific geographic coordinates, allowing for the identification of hotspots based on locations. Asset management records were integrated into the dataset to enrich the historical outage data analysis and enhance the predictive models. This integration provided essential insights into the functional locations of equipment, equipment age, and various other attributes [9]. The asset management records, which were sourced from local utility companies and infrastructure management systems, offered detailed information about the equipment deployed in the eastern region of the City of Cape Town municipality. The integration process involved matching equipment IDs (identities) from the historical outage dataset with corresponding entries in the asset management records. This allowed for the extraction of essential attributes, including the equipment's functional location within the network, its age, condition, manufacturer details, and technical specifications.

3.2. Load shedding data

The study utilized historical data on load shedding, providing insights into the electricity supply disruptions experienced during the study period. By incorporating load-shedding data, the research aimed to identify patterns and correlations between load-shedding events and equipment failures, contributing to improved outage prediction and maintenance strategies [10]. To achieve this integration, the load-shedding data was thoughtfully merged with outage data, aligning it with specific load-shedding areas and corresponding substations. This meticulous process allowed for a comprehensive examination of how load-shedding events impacted equipment failures within specific geographical regions. By analyzing the timing and extent of load shedding concerning outage occurrences, the study sought to reveal causal relationships and further refine predictive models for increased accuracy.

3.3. Weather data

A set of pertinent weather parameters were selected to be integrated into the analysis. These parameters were chosen due to their potential impact on electrical equipment and their influence on the onset of outages [11]. The selected weather parameters include maximum temperature, minimum temperature, average temperature, humidity, dew point, UV index, precipitation, and wind speed. Each of these parameters were considered for their unique effects on equipment performance and its relevance to predicting equipment outages. Extreme temperature variations, humidity levels, and other weather conditions can result in undesirable equipment behavior or outages, making them valuable model features [12].

Weather data was meticulously obtained from various external sources to cover the area of interest. Specifically, the research focused on the eastern region of the City of Cape Town municipality. The weather records were obtained for the period spanning from January 2018 to May 2023. The data was sourced from regional meteorological agencies, weather monitoring stations, and publicly accessible weather databases. These datasets were instrumental in capturing the local weather conditions surrounding the electrical substations.

The integration of weather data was facilitated through thorough preprocessing. This step aimed to ensure that the weather data was compatible with the historical outage dataset and was consistent in format and quality. The historical weather records were harmoniously merged with the historical outage dataset. This integration was orchestrated based on the timestamp of each data point, effectively aligning weather conditions with the corresponding outage events. The resulting unified dataset represented an amalgamation of outage records and associated weather conditions, creating a holistic view of the relationship between weather parameters and equipment failures. This integrated dataset, enriched with a plethora of meteorological information, served as the foundational input for the development of the machine learning models. It allowed these models to harness the predictive potential encapsulated within weather parameters and facilitated the identification of intricate patterns and correlations between prevailing weather conditions and equipment performance. The fused dataset formed the heart of the research's analytical framework and was instrumental in unravelling the complex interplay between weather and equipment outages [12].

3.4. Addressing inaccurate data components

The initial dataset undergoes a filtering process to eliminate undesirable data points. These undesirable values encompass outliers, missing data points, and duplicates. Outliers refer to those data points within historical outage data that exhibit significant deviations from the rest of the dataset, primarily resulting from recording errors [13]. Outliers are removed from the dataset, as their presence would not significantly influence the overall distribution. A mean imputation technique was used for missing values in the equipment data.

3.5. Preparation and transformation of features and targets

The selected features underwent data preprocessing to ensure uniformity, consistency, and suitability for machine learning models. This included standardization (scaling features to a common range), normalization (adjusting data to a standard distribution), and handling of missing values. Since the dataset contained categorical features, encoding techniques were applied to convert these variables into numerical format. The target variables were encoded into a format suitable for machine learning. One-hot encoding was applied to transform the target into categorical variables.

4. Numerical experiment

In this project, two separate machine learning models were developed due to the lack of data attributes for cables and overhead lines, particularly age and manufacturer information. The primary motivation behind creating separate models was to avoid the challenges associated with missing data. By splitting the project into two models, one focusing on equipment and the other on cables and lines, it was possible to effectively utilize the available data without the need to predict or fill in missing values for attributes like age and manufacturer. This approach ensures that the models are trained on complete and reliable datasets, leading to more precise predictions for equipment and cable/line scenarios.

A 70–30 split ratio is used for both models, allocating 70% of the data for training and reserving 30% for testing in each case. The division of the data was performed using a randomized splitting technique to maintain the diversity and representativeness of both the training and testing datasets. Specifically, the splitting process was carried out using the 'train_test_split' function from the 'sklearn.model_selection' library. This function shuffles the dataset randomly before splitting, which helps reduce bias and variance in the model's performance.

To build the models, a random forest classifier was employed. The diversification of information for the random forest model was achieved through two primary mechanisms: bootstrap sampling and feature randomness. Bootstrap sampling (bagging) ensures that each decision tree in the random forest is trained on a different subset of the training data, created by randomly sampling with replacement. This method helps in reducing overfitting and improving the model's generalization ability. Additionally, feature randomness (random subspace method) involves selecting a random subset of features for each split in the decision trees, which ensures that different trees focus on different aspects of the data, thereby increasing the model's robustness and accuracy.

Before hyperparameters tuning, the models displayed signs of severe overfitting, with a training loss of zero and perfect training accuracy, indicating an excessively optimistic fit to the training data [13]. However, this was accompanied by a drop in performance on unseen data, as evidenced by the lower validation accuracy. To address these challenges the hyperparameters in Table 2 were selected for the random forest classifier (RFC). The number of trees, “n_estimators”, was set at 100, providing a balance between computational efficiency and model accuracy, thus ensuring the ensemble is robust enough to generalize well beyond the training data. The maximum depth of each tree, “Max_depth”, was capped at 10 to prevent the model from fitting excessively to noise in the training data, which helps reduce overfitting while still allowing the model to capture underlying data patterns. The “Min_samples_split” was set to 2, allowing the trees to split even at low sample sizes, which increases the model’s sensitivity to nuances in the data but requires careful monitoring to prevent overfitting. Lastly, “Min_samples_leaf” was set at a minimum of 1, ensuring that the trees capture very fine details to improve model granularity, though this setting also raises the potential for overfitting, which was mitigated through rigorous cross-validation.

Table 2. Hyperparameters of the RFC

Hyperparameters	**Value**
n_estimators	100
Max_depth	10
Min_samples_split	2
Min_sample leaf	1

The implementation of these hyperparameters effectively mitigated the overfitting issue, as indicated by the increased training loss and the slight decrease in training accuracy, without compromising the validation performance, which remained consistent as shown in Table 6.

Additionally, to address the challenge of class imbalance particularly in the equipment model, the synthetic minority over-sampling technique (SMOTE) was applied. This intervention improved the balance of the training dataset, as reflected in the enhanced validation accuracy from Table 3, which shows metrics before and after hyperparameter tuning and the application of SMOTE. This approach not only improved the model’s ability to generalize but also its precision in handling under-represented classes.

Table 3. Equipment model training and validation metrics

Metrics	**Initial**	**With hyperparameter tuning**	**With SMOTE**
Training loss	0.0000	0.0424	0.0449
Validation loss	0.1329	0.1329	0.1272
Training accuracy	1.0000	0.9576	0.9551
Validation accuracy	0.8671	0.8671	0.8728

The equipment initial model showed strong performance in some classes while struggling with others, particularly those with limited representation in the dataset (Class 5, 9, 10, and 12). SMOTE (synthetic minority over-sampling technique) was applied for the training set to rectify the class imbalance, resulting in a more balanced dataset. SMOTE over-sampling has also improved the validation accuracy, as indicated in Table 3 and Table 4 results.

Table 4. Cable and Line model training and validation metrics

Metrics	Normal	With hyperparameter tuning
Training loss	0.0000	0.0390
Validation loss	0.1553	0.1667
Training accuracy	1.0000	0.9610
Validation accuracy	0.8447	0.8333

Figure 2 and Fig. 3 present the metrics graphs before the models' final tunings. Figure 4 and Fig. 5 present the metrics graphs after the models' final tunings.

The chosen cross-validation method is 5-fold cross-validation, as indicated by the "$v = 5$" parameter. In 5-fold cross-validation, the dataset is divided into five roughly equal-sized subsets [4, 8]. The model is trained and evaluated five times, each using a different fold as the test set, while the remaining four folds are used for training. This process is repeated five times to assess the model's performance thoroughly. To be more specific:

$$CV_k = \sum_{k=1}^{K} \frac{n_k}{n} l_{Ck}\left(g_{\tau_{-k}}\right), \tag{1}$$

$$CV_k = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} \text{Loss}\left(g_{\tau_{-k}}\left(x_i\right), y_1\right), \tag{2}$$

$$CV_k = \frac{1}{n} \sum_{k=1}^{n} \text{Loss}\left(g_{\tau_{-k}}\left(x_i\right), y_1\right), \tag{3}$$

where the function of $K : \{1, \ldots, n\} \longmapsto 1, \ldots < K]$ indicates to which of the K folds each of the n observations belongs. As the average is taken over varying training sets, $\mathcal{T}_{-k}$, it estimates the expected generalization risk ($\mathbb{E}\ell g_{\mathcal{T}}$), rather than the generalization risk ($g_{\mathcal{T}}$) for the particular training set τ [13]. The cross-validation scores obtained for the equipment model are (0.83478261, 0.83478261, 0.82608696, 0.85217391, and 0.8245614) with an average of 0.83, which indicates the model's performance in each of the five iterations of the cross-validation process. The high scores indicate the model can predict equipment failures based on the provided features. The consistency of the scores across different folds suggests that the model is not overfitting or underfitting the data but is making reliable predictions. For cables and overhead lines, the model produced validation scores of (0.77840909, 0.73863636, 0.80113636, 0.8125, and 0.73714286) with an average of 0.77, which suggests that the model is reasonably robust and can perform consistently across different data splits, which is a positive outcome considering the missing data features like age and make.

Identifying a subset of critical features essential for precise prediction in a random forest can be challenging. The feature importance measure intends to address this issue. The feature importance measure is:

$$I_{RF}\left(x_j\right) = \frac{1}{B}\sum_{b=1}^{B} I_{T_b}\left(x_j\right), \quad 1 \le j \le p. \tag{4}$$

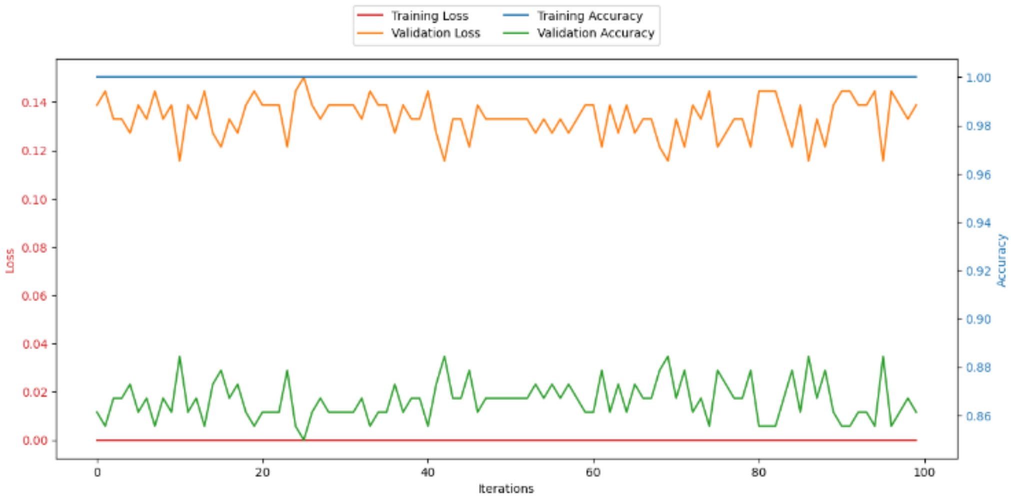

Fig. 2. Equipment model training and validation metrics without hyperparameters tuning, 5-fold cross-validation and SMOTE

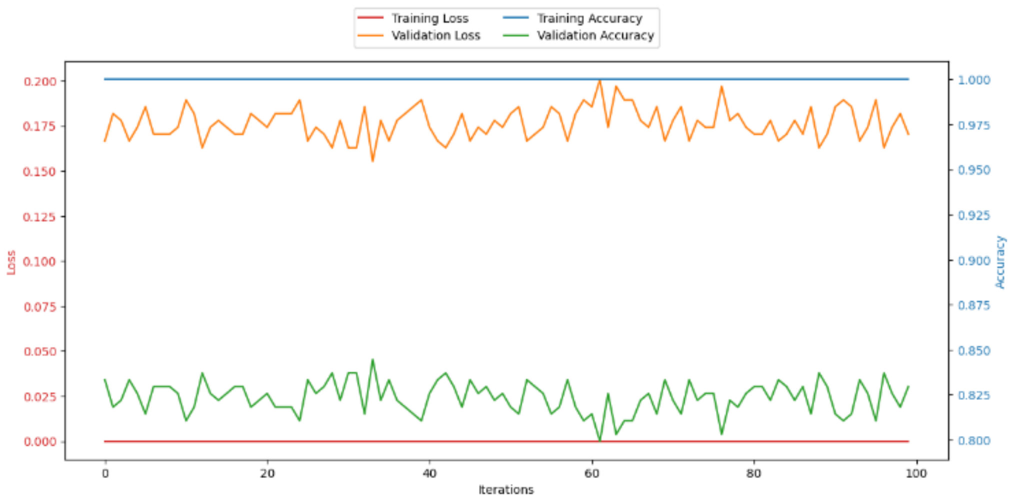

Fig. 3. Cable and Line model training and validation metrics without hyperparameters tuning and 5-fold cross-validation

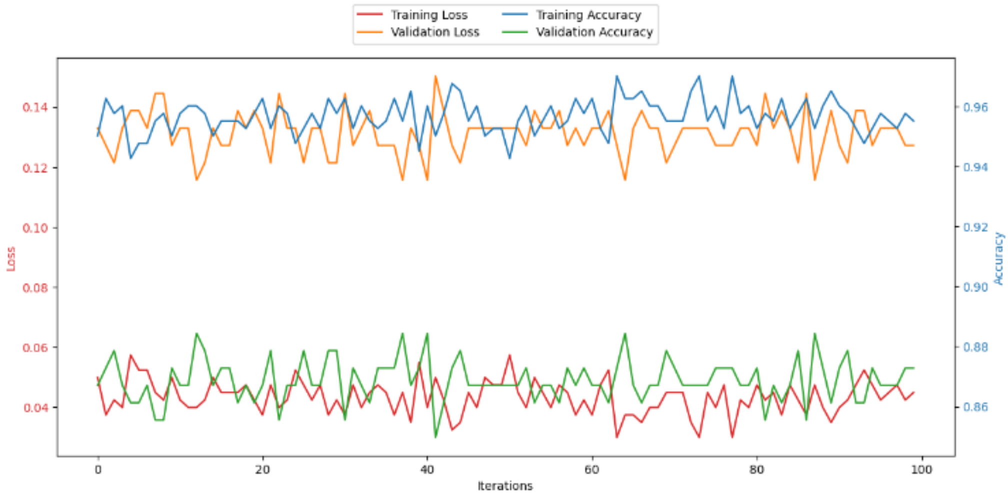

Fig. 4. Equipment model train and validation metrics. Hyperparameters tuning, 5-fold cross-validation, and SMOTE over-sampling were applied

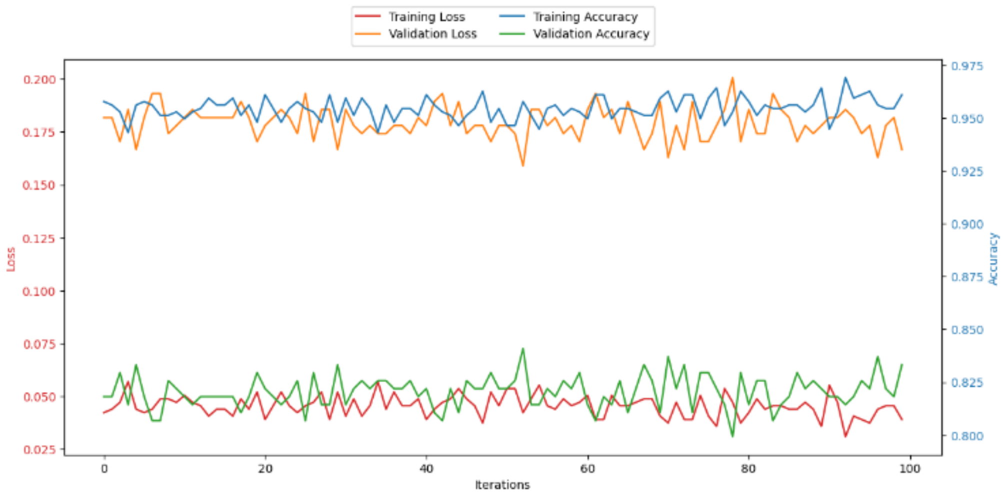

Fig. 5. Cable and Line model train and validation metrics. Hyperparameters tuning and 5-fold validation were applied

The feature importance of the equipment model in Fig. 6 indicates that equipment category, manufacturer type, and age are the most significant. Other moderately important features are wind speed, equipment, soil, and age. The weather features exhibit a wide range of importance with varying individual impacts, with some features having a more substantial influence on the model's decision-making process than others.

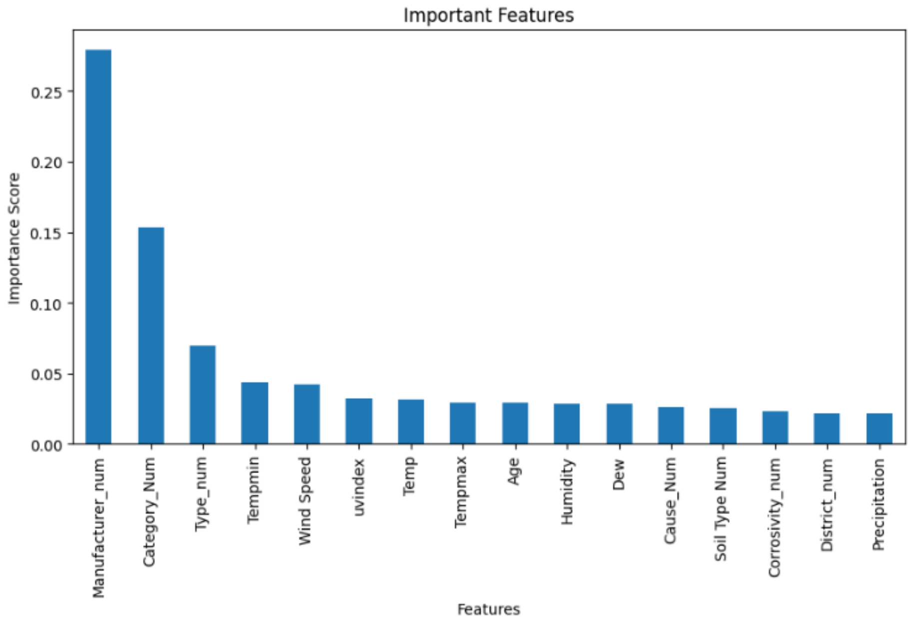

Fig. 6. Feature importance in the equipment failure prediction model

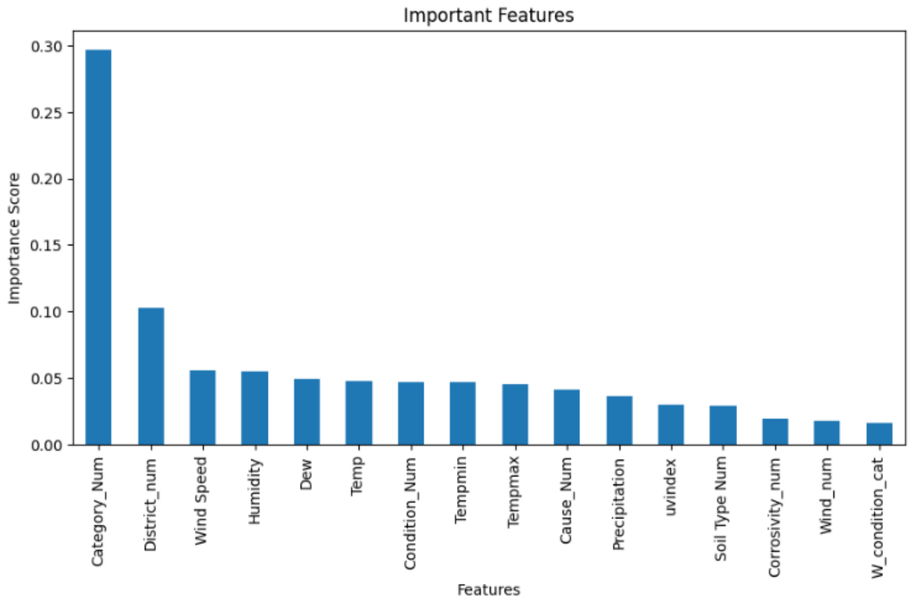

Fig. 7. Feature importance in the cable and overhead line failure prediction model

The feature importance analysis for the cable and overhead line model in Fig. 7 indicates that the most critical factors influencing its predictions include equipment category, humidity, dew point, temperature, and district. These elements are suggested to have a significant impact on cable and overhead line issues. Additionally, features such as maximum temperature, wind speed, minimum temperature, cause of outage, and UV index are considered moderately important, underscoring their role in predicting these problems to a certain degree. The analysis reveals that various weather-related features exhibit different levels of importance; for instance, humidity and dew point are particularly influential, whereas the UV index, though less prominent, still plays a relevant role in the model's decision-making process.

The hyperparameters collectively configure the RFC. The choice of number of trees, depth of each tree, minimum sample splits and leaf impact how the model captures patterns in the data. The application of hyperparameters had varying effects on both models. For the Equipment model, regularization successfully mitigated overfitting, leading to a more balanced model. Although some classes improved recall, the overall performance, measured by the weighted average F1-score, remained unchanged at 0.81.

On the other hand, the Cable and Line model demonstrated notable enhancements after hyperparameters tuning. Precision, recall, and F1-scores improved for most fault classes, resulting in a higher weighted average F1-score of 0.82. The final setup for hyperparameters is listed in Table 5.

Table 5. Hyperparameter tuning of the RFC

Hyperparameters	**Value**
n_estimators	100
Max_depth	10
Min_samples_split	2
Min_sample leaf	1

The dataset contained 1521 samples from 1 January 2018 to 30 May 2023. The missing parameters were initialized to zero to ensure the technical coherence of the data. The data contained 86 unlabeled parameters where an outage occurred, but the equipment category was not labelled. This data is labelled manually as unknown equipment, which did not negatively affect the model's accuracy.

5. Results

The model's results performance was evaluated with accuracy, precision, recall and F-1 score using both micro average and macro average.

5.1. Model performance in multi-class classification

In the initial equipment model, performance metrics displayed a mixed picture. It demonstrated high precision and recall for certain classes, notably classes 8 and 11, but it struggled significantly with others, including classes 5, 9, 10, and 12. This imbalance was evident in the F1-scores and highlighted the need for model improvement. The accuracy appeared reasonably high at 87%, mainly due to the strong performance in a few classes. After incorporating regularization parameter tuning and SMOTE, the model's performance remained largely consistent, with minor improvements in certain metrics. While some classes saw increased recall, particularly class 5, the model still faced challenges in achieving a balanced performance across all classes.

Table 6. Equipment model random forest classifier results. RFC Xht represents hyperparameters tuning to improve the model

Metrics	RFC	RFC Xht	RFC SMOTE
Accuracy	87%	87%	83%
Precision macro average	53%	67%	66%
Precision weighted average	78%	81%	84%
Recall macro average	48%	48%	59%
Recall weighted average	87%	87%	83%
F1-score macro average	48%	49%	61%
F1-score weighted average	81%	81%	83%

Table 7. Cable and Line model random forest classifier results. RFC Xht represents hyperparameters tuning to improve the model

Metrics	RFC	RFC Xht
Accuracy	79%	83%
Precision macro average	76%	87%
Precision weighted average	79%	84%
Recall macro average	76%	81%
Recall weighted average	79%	83%
F1-score macro average	76%	83%
F1-score weighted average	79%	82%

In contrast, the Cables and Lines model exhibited more balanced results from the outset. In its initial state, the model showed decent precision, recall, and F1-scores for most classes, with an overall accuracy of 79%. The macro and weighted average F1-scores were also reasonable at 76% and 79%, respectively. After fine-tuning with regularization, the model's performance saw notable improvements. Precision and recall for several classes, such as classes 3 and 11, were enhanced, resulting in higher F1-scores. The model's overall accuracy increased to 83%, and both macro and weighted average F1-scores improved to 83% and 82%, respectively. These enhancements suggested that the model for Cables and Lines prediction benefited significantly from the applied modifications.

While both models initially exhibited some performance issues, the Cables and Lines model showed more balanced results and responded positively to regularization and SMOTE. On the other hand, despite slight improvements, the equipment outage prediction model continued to face challenges in addressing class imbalances and achieving consistent performance across all categories. Further refinements and possibly alternative modelling approaches may be necessary to improve its overall effectiveness.

Performance is evaluated with accuracy, precision, recall and F1 score using macro and weighted averages. The first model represents the system's initial state, while the second model incorporates hyperparameter tuning for optimization. The initial model, configured with default hyperparameters, exhibited varying precision scores across classes ranging from 0.00 to 1.00. Certain classes, such as 6, 10, and 12, presented challenges in correctly identifying positive instances. However, recall scores were generally high, approaching 1.00, indicating a proficient ability to recognize relevant instances. The F1-score, reflecting a harmonic mean of precision and recall, also tended to be high, suggesting a balanced trade-off between these performance metrics.

5.2. Model probability

The class probabilities generated by the equipment model provide valuable insights into its prediction process for a specific instance (Fig. 8). This analysis observes a wide range of probabilities assigned to different classes. Class 11 has the highest probability of 0.5150, indicating the model's strong confidence in this classification. Conversely, Class 9 receives a probability of 0.0000, signalling absolute certainty that the instance does not belong to this class. Class 10, with a probability of 0.2200, is the second most likely classification, suggesting moderate confidence. Classes 5, 8, 12, and 14 receive lower probabilities, implying lower confidence. Class 8, for instance, has a probability of 0.0350, indicating a slightly increased likelihood but not a definitive classification. These results highlight the model's varying confidence levels in different classes and provide valuable information for understanding its decision-making process for this particular instance.

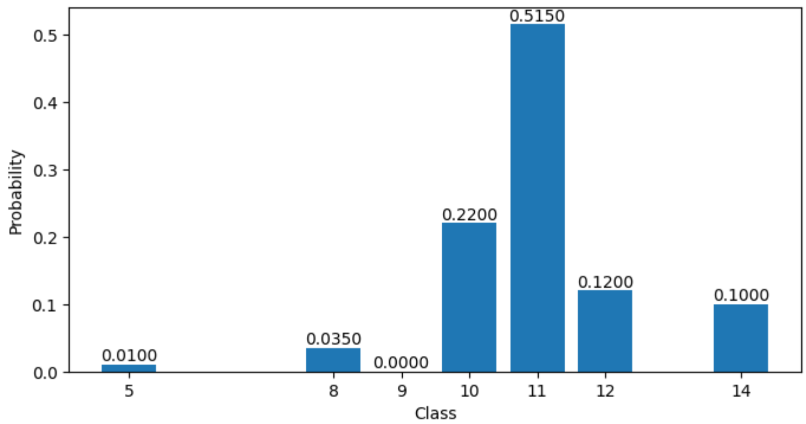

Fig. 8. Equipment model class probabilities for the test instance

The class probabilities generated by the model for the Cables and Lines classification reveal the model's level of confidence in assigning the specific instance to different classes (Fig. 9). In this analysis, we observe a spectrum of probabilities, each reflecting a distinct degree of certainty. Class 11 and Class 13 emerge as the most likely classifications, with probabilities of 0.4090 and 0.4475, respectively. These high probabilities underscore the model's robust confidence in these two classes for this instance. On the other hand, Classes 2 and 3 receive notably lower probabilities, at 0.0867 and 0.0569, respectively. These lower probabilities indicate the model's hesitancy and uncertainty regarding these classifications. While not ruling out the possibility entirely, the model is cautious about assigning the instance to these classes. These class probabilities provide valuable insights into the model's decision-making process, showcasing its varying confidence levels and highlighting the most likely classifications for this specific instance in the Cables and Lines context.

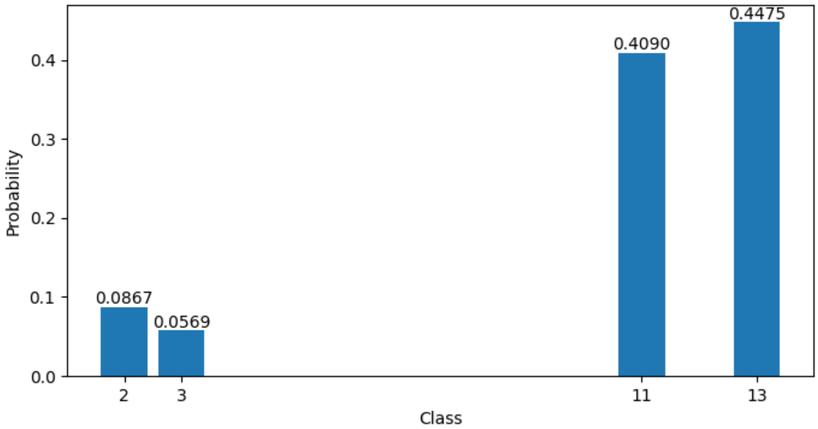

Fig. 9. Cable and Lines model class probabilities for the test instance

5.3. Confusion matrix

The confusion matrix shown in Fig. 10 and Fig. 11 offers a comprehensive view of the equipment classification model's performance across various classes. It provides a breakdown of correct and incorrect predictions, which is essential for assessing the model's effectiveness. The model struggles with some classes, particularly Classes 3 and 5, where misclassifications occur, as evidenced by non-zero values outside the diagonal. On the other hand, Classes 1 and 2 show relatively good performance, with the majority of samples correctly classified. However, the presence of class imbalance is still evident in Class 4, where only a few samples are present. In the normalized confusion matrix, some classes exhibit perfect accuracy, as evidenced by diagonal entries with values of 1.00, indicating that all samples in those classes were correctly classified.

On the other hand, the Cable and Line model demonstrates proficiency in classifying certain categories, such as Class 1 and Class 4, as indicated by the high number of true positives. However, challenges arise in correctly identifying instances for other classes, such as Class 2 and Class 3, where false positives and false negatives occur. The absence of true negatives suggests that the model didn't correctly identify any instances as not belonging to any of the classes. The confusion matrix and the normalized confusion matrix are shown in Fig. 12 and Fig. 13.

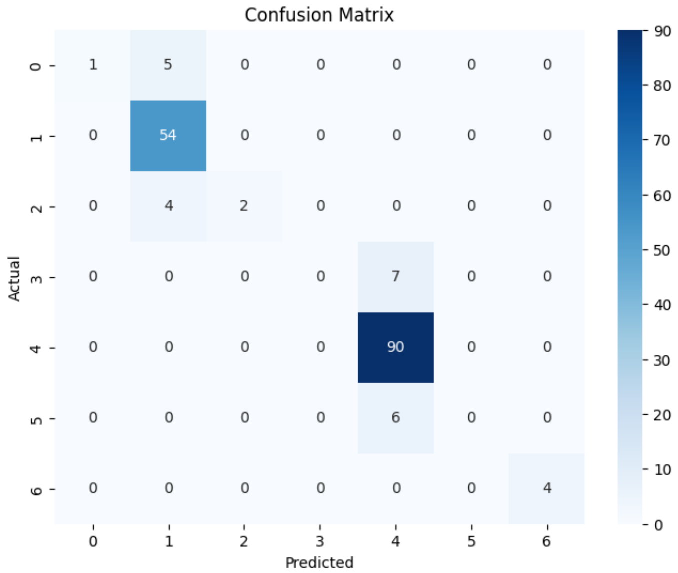

Fig. 10. Equipment model confusion matrix. Each cell represents the performance of a classification model by showing the counts of correctly and incorrectly classified instances across different categories

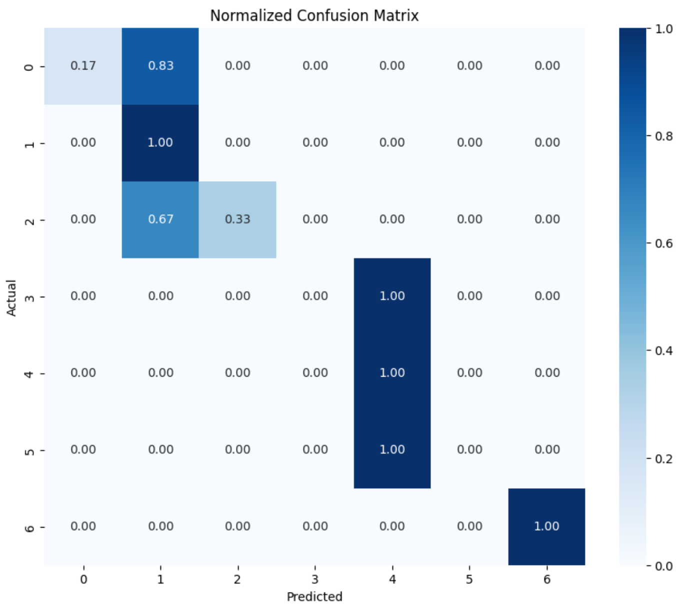

Fig. 11. Equipment model normalized confusion matrix. The cells represent the percentages or proportions of correct and incorrect classifications in each category

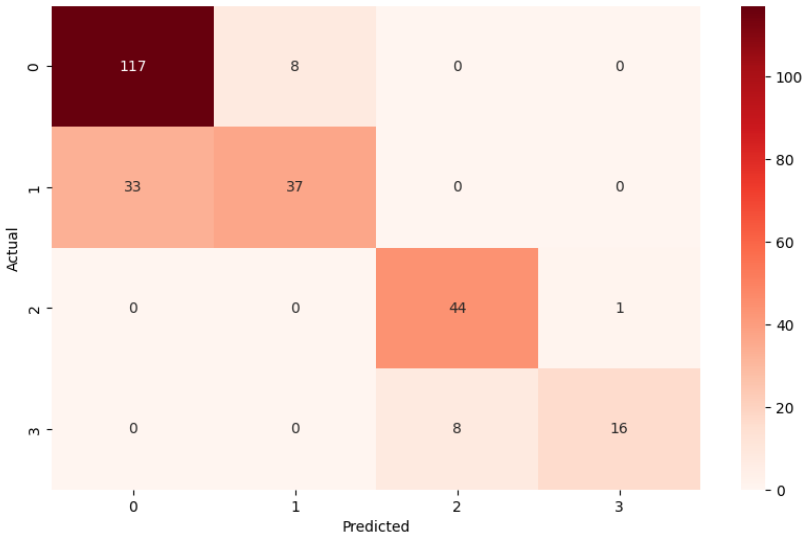

Fig. 12. Cable and Line model confusion matrix

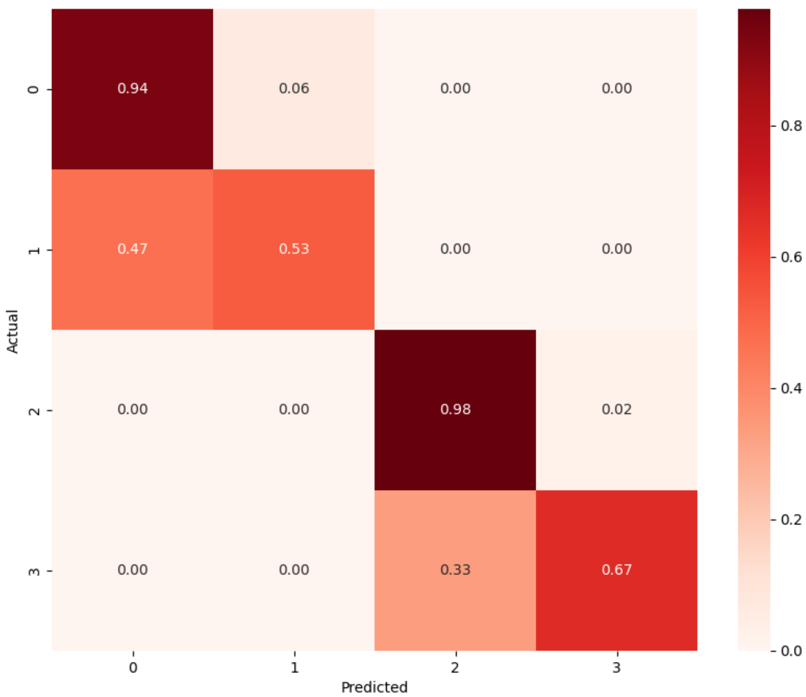

Fig. 13. Cable and Line model normalized confusion matrix

6. Conclusions

The project has made several significant contributions to predictive maintenance and reliability engineering. One notable accomplishment was effectively handling missing data attributes, particularly age and manufacturer information, by developing separate machine-learning models for equipment and cables/lines. This strategic approach allowed for the utilization of available data without the complexities of imputation or predicting missing values. Furthermore, the project successfully mitigated class imbalance issues, a common challenge in equipment failure prediction, by applying the synthetic minority over-sampling technique (SMOTE), resulting in more balanced training datasets and improved model performance.

The Cables and Lines model exhibited more balanced and robust performance from the outset, showing decent precision, recall, and F1-scores for most classes, with an overall accuracy of 79%. After fine-tuning with regularization, the model's performance saw notable improvements, with its accuracy increasing to 83%, and both macro and weighted average F1-scores improving to 83% and 82%, respectively.

In contrast, the equipment failure prediction model faced persistent challenges in achieving balanced performance across different equipment classes. Despite incremental improvements following regularization and SMOTE application, the model continued to struggle with class imbalances, indicating the need for further refinements and alternative modeling approaches to enhance its effectiveness.

Further refinements and alternative modelling approaches should be explored to enhance the effectiveness of the equipment failure prediction model. Potential future research directions could include the integration of more advanced techniques such as ensemble learning, deep learning models, or the application of domain-specific knowledge to better handle class imbalances and improve model interpretability. Specific examples of alternative approaches include exploring gradient boosting methods, employing neural networks tailored to the specific characteristics of equipment data, and leveraging transfer learning to apply knowledge from similar domains to improve prediction accuracy and reliability.

Acknowledgements

I would like to extend my sincere appreciation to the City of Cape Town municipality for their invaluable support and authorization to access and utilize their data, which was pivotal in the execution of this research project. I would like to express my deepest gratitude to my supervisor, Professor M. Sumbanyambe. His unwavering guidance, mentorship, and dedication to my growth as a researcher have been instrumental in the success of this project. I am genuinely thankful for his willingness to devote time to my academic and professional development.

References

[1] Asaridis P., Molinari D., Ballio F., *A conceptual model for the estimation of flood damage to power grids* (2021), DOI: 10.5194/egusphere-egu21-2721.

[2] Nduhuura P., Garschagen M., Zerga A., *Impacts of Electricity Outages in Urban Households in Developing Countries: A Case of Accra, Ghana*, Energies, vol. 14, no. 12, 3676 (2021), DOI: 10.3390/en14123676.

[3] Onaolapo A., Carpanen R., Dorrell D., Ojo E., *A comparative assessment of conventional and artificial neural networks methods for electricity outage forecasting*, Energies, vol. 15, no. 2, 511 (2022), DOI: 10.3390/en15020511.

[4] O'Fallon C., Gopstein A., *Quantifying operational resilience benefits of the smart grid* (2021), DOI: 10.6028/nist.tn.2137.

[5] Kiran D., *Failure Modes and Effects Analysis. Total Quality Management*, pp. 373–389 (2017), DOI: 10.1016/B978-0-12-811035-5.00026-X.

[6] Tsioumpri E., Stephen B., McArthur S., *Weather Related Fault Prediction in Minimally Monitored Distribution Networks*, Energies, vol. 14, no. 8, 2053 (2021), DOI: 10.3390/en14082053.

[7] Fakih A., Ghazalian P., Ghazzawi N., *The Effects of Power Outages on the Performance of Manufacturing Firms in the Mena Region*, Review of Middle East Economics and Finance, vol. 16, no. 3 (2020), DOI: 10.1515/rmeef-2020-0011.

[8] Luo G., Huang F., Yang Y., Zhang B., Chen Z., Peng W., *Intelligent aided decision-making method for power grid outage maintenance plan based on machine learning* (2023), DOI: 10.1117/12.2686169.

[9] Service E.A., *Eskom*, March 2020 [online], accessed December 2023.

[10] Evans J., *Daily Maverick*, 27 June 2022, available at: https://www.dailymaverick.co.za/article/2022-06-27-pushing-the-limits-why-load-shedding-puts-even-more-pressure-on-an-ageing-electrical-system/, accessed 12 March 2023.

[11] Berk A., *Dynamic Modeling of Power Outages Caused by Thunderstorms* (2020), DOI: 10.3390/forecast2020008.

[12] Godemel F., *Powermag*, Power, 11 November 2022 [online], available at: https://www.powermag.com/blog/power-outages-due-to-extreme-weather-must-become-a-thing-of-the-past/, accessed 12 March 2023.

[13] Géron A., *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*, United States of America, O'Reilly Media, ISBN: 9781492032649 (2019).