World Journal of Advanced Research and Reviews

eISSN: 2581-9615 CODEN (USA): WJARAI

Cross Ref DOI: 10.30574/wjarr

Journal homepage: https://wjarr.com/

(Research Article)

Modeling of faults in the CEB electrical transmission network by approaches: KNN, Random Forests, logistic regression, SVM, ANN and gradient boosting of supervised learning

Sadissou KARIMOUN IBRAHIM [1], Komla Kpomonè APALOO BARA [1, 2, 3, *] and Yao BOKOVI [1, 2, 3]

[1] Regional Center of Excellence for Electricity Control (CERME), University of Lomé, Lomé, Togo.
[2] Department of Electrical Engineering, Polytechnic School of Lomé (EPL), University of Lomé, Togo.
[3] Engineering Sciences Research Laboratory (LARSI), University of Lomé, Togo.

World Journal of Advanced Research and Reviews, 2024, 24(03), 2099-2115

Publication history: Received on 03 November 2024; revised on 18 December 2024; accepted on 20 December 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.24.3.3802

Abstract

The work accumulated in this article presents the results of learning the faults that affect the CEB network. The objective is to predict failures in order to prevent these faults from creating interruptions. The network operating data from 2008 to 2015 are used as materials. The algorithms: SVM, KNN, Random Forest, Gradient Boosting, ANN and Logistic Regression were used as methods to create the models. The results are subjected to evaluation criteria namely: the confusion matrix, the area under the ROC curve and the scores (Accuracy, F1 Score, Precision and recall). A characterization of the faults is carried out. The results of the characterization reveal that there are 19 faults and the most recurrent is the short circuit, which appeared 947 times out of 2427 during the study period. The modeling results are perfect. The True Positives of the confusion matrices are greater than 450 out of 497, for the classes. Some are better than others. The unfavorable is obtained through the KNN with AUC=0.761. Its score, (Accuracy=0.955; F1 Score = 0.957; Precision=0.958; Recall=0.955), confirms this observation. The AUC=0.664, remains even more unfavorable with the SVM modeling but its score, (Accuracy=0.989; F1 Score = 0.988; Precision=0.988; Recall=0.989), exceeds that of KNN. Moreover, for the other models, their AUC exceeds 80% with the more perfect logistic regression giving: AUC=0.991; Accuracy=0.991; F1 Score = 0.991; Precision=0.991; Recall=0.991. These results confirm that, even very random and of various causes, we can predict the defects in the CEB network. However, it is necessary to use more recent data in order to apply these results in future operations.

Keywords: ANN; Defects; Gradient boosting; KNN; Logistic regression; Modeling; Power network; Random forest; SVM

1. Introduction

Human habits in today's world lead us to understand that after the industrial revolution, we are in the digital revolution. We find ourselves in a world where almost everything is going digital except the transport of users and goods; agriculture is no less. In this context, we can distinguish the circulation of money, communications, conferences and even the vast majority of acts of violence that are no longer carried out by human contact but by devices. Among the devices that allow most of the aforementioned operations to be carried out, we can list: mobile phones, computers, servers, drones, remote controls, etc. Whether these devices are fixed or mobile, the heart of their operation is electricity. Several primary energy sources are used to produce this electrical energy [1, 2]. These are fossil sources (natural gas, coal, oil), polluting nature by their release of greenhouse gases [3, 4, 5]. Technology has evolved so much since the 2022s that we are turning our eyes towards small modular nuclear reactors which have very interesting yields. That being said, international policies are currently encouraging production based on renewable and clean sources (sun,

[*] Corresponding author: APALOO BARA Komla Kpomonè

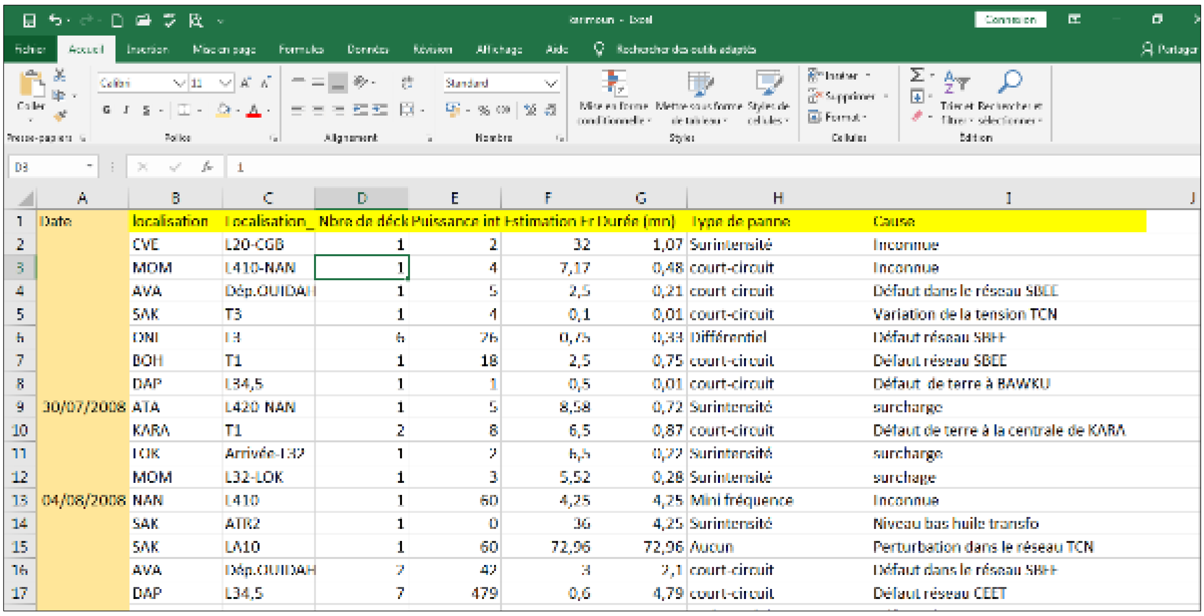
wind, geothermal) [6, 7]. The transformation of these sources into electricity is done in power plants, namely: hydroelectric, thermal flame or nuclear, geothermal, wind, solar photovoltaic [8, 9, 10].

However, these power plants are not close to places of mass consumption. We thus find the need to install the networks for the transport and distribution of electrical energy [11, 12]. Despite the consideration of several protection devices allowing the sustainability of the supply of electrical energy, we always come across some faults in places in the networks. This sometimes causes short or long interruptions. In order to remedy these problems, we aim through this work to use modern methods to predict failures. The literature shows that several works using different methods (modeling, optimization and statistical characterization), have taken into account the prediction and resolution of faults in the networks, [13, 14, 15, 16]. They have repeatedly used artificial intelligence. In artificial intelligence, there is supervised learning, unsupervised learning, [17], and deep learning [18]. These types of learning make it possible to predict random situations from the observations made.

For this work, we choose supervised learning. It includes several algorithms among which we can list: SVM, K-Nearest Neighbors, Random Forest, Gradient Boosting, Artificial Neural Network, Logistic Regression. They will allow us to learn the failures noted in the CEB transmission network. The objective is to predict failures in order to prevent faults from creating interruptions in the supply of electricity. To achieve this, we will use operating data related to the recording and recurrence of faults in the network of the Benin Electric Community (CEB). This is a company responsible for transporting electrical energy to cover Togo and Benin; which are countries in humid and coastal West Africa. In Togo, energy distribution is the responsibility of the Togo Electric Power Company (CEET) and the Beninese Electric Power Company (SBEE) is responsible for that of Benin [19, 20, 21]. The aim of this work is to learn the frequency of failures in these networks and to submit the test results, obtained with the algorithms used, to some performance evaluation metrics in order to judge the effectiveness of the models. Among the metrics, we retain: the confusion matrix, the ROC curve and the scores (Accuracy, F1 Score, Precision and recall). The results will make it possible to anticipate through preventive maintenance, the defects in order to maintain the sustainability of the electricity supply in the CEB network.

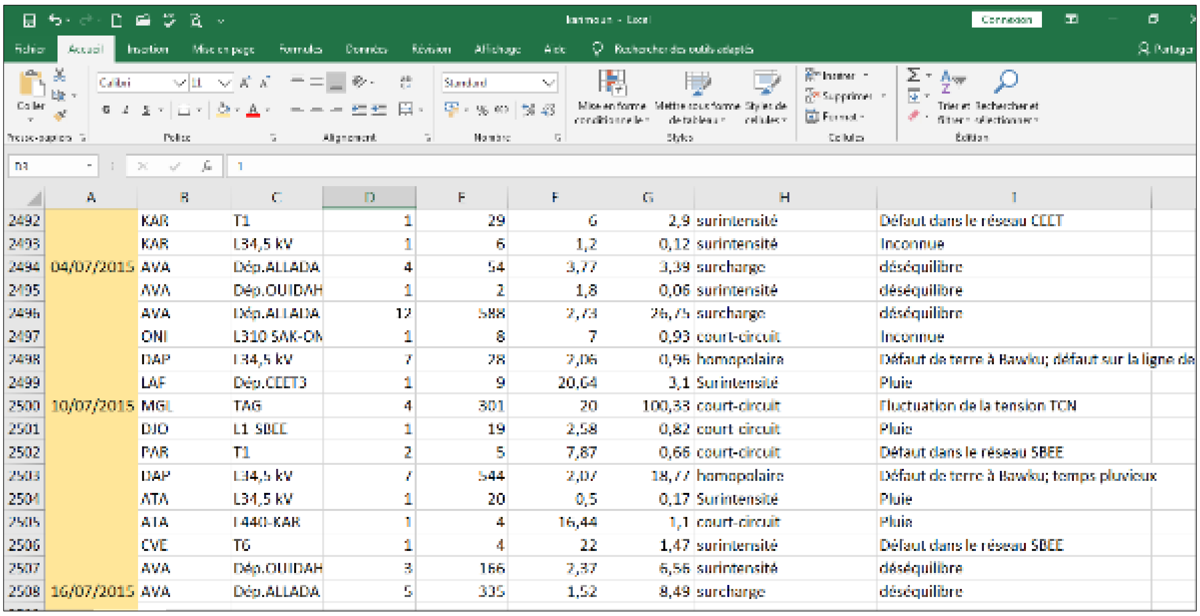
2. Material and methods

The CEB networks are made by overhead and underground lines in places. We only focus on the overhead lines in this work. It starts in Ghana and ends in Benin, passing particularly through the Nangbéto hydroelectric power plant, with a capacity of 75 MW which also makes it possible to manage peak hours (hours at which peaks in power consumption are high). For its management, there is a dispatching that controls all departures and arrivals. Interruptions commonly occur for two reasons. In the first case, the inability to meet demand due to insufficient production or in the event of aberrant demand. The second case comes when an unexpected failure occurs on the network. In this case, the causes are often technical failures, human errors, climatic conditions. It is the triggering of protection devices that often signals interruptions in the supply of electrical energy to customers. In this case, the operators report hierarchically who decide to move on to restoration or do maintenance before resumption. Following this, weekly recordings are made as an operating report, [22]. In this work, we exploit these data collected from 2008 to 2015 on the network of which Figure 1 shows the beginning and Figure 2 the end of the Excel recording sheet.



	A	B	C	D	E	F	G	H	I
1	Date	localisation	Localisation_	Nbre de déck	Puissance int	Estimation Er	Durée (mn)	Type de panne	Cause
2		CVE	L20-CGB	1	2	32	1,07	Surintensité	Inconnue
3		MOM	L410-NAN	1	4	7,17	0,48	court-circuit	Inconnue
4		AVA	Dép.OUIDAH	1	5	2,5	0,21	court circuit	Défaut dans le réseau SBEE
5		SAK	T3	1	4	0,1	0,01	court-circuit	Variation de la tension TCN
6		ONI	L3	6	26	0,75	0,33	Différentiel	Défaut réseau SBEE
7		BOH	T1	1	18	2,5	0,75	court-circuit	Défaut réseau SBEE
8		DAP	L34,5	1	1	0,5	0,01	court-circuit	Défaut de terre à BAWKU
9	30/07/2008	ATA	L420 NAN	1	5	8,58	0,72	Surintensité	surcharge
10		KARA	T1	2	8	6,5	0,87	court-circuit	Défaut de terre à la centrale de KARA
11		LOK	Arrivée-L32	1	2	6,5	0,22	Surintensité	surcharge
12		MOM	L32-LOK	1	3	5,52	0,28	Surintensité	surchage
13	04/08/2008	NAN	L410	1	60	4,25	4,25	Mini fréquence	Inconnue
14		SAK	ATR2	1	0	36	4,25	Surintensité	Niveau bas huile transfo
15		SAK	LA10	1	60	72,96	72,96	Aucun	Perturbation dans le réseau TCN
16		AVA	Dép.OUIDAH	2	42	3	2,1	court-circuit	Défaut dans le réseau SBEE
17		DAP	L34,5	7	479	0,6	4,79	court-circuit	Défaut réseau CEET

Figure 1 Start of the CEB operating data collection pages



	A	B	C	D	E	F	G	H	I
2492		KAR	T1	1	29	6	2,9	surintensité	Défaut dans le réseau CEET
2493		KAR	L34,5 kV	1	6	1,2	0,12	surintensité	Inconnue
2494	04/07/2015	AVA	Dép.ALLADA	4	54	3,77	3,39	surcharge	déséquilibre
2495		AVA	Dép.OUIDAH	1	2	1,8	0,06	surintensité	déséquilibre
2496		AVA	Dép.ALLADA	12	588	2,73	26,75	surcharge	déséquilibre
2497		ONI	L310 SAK-ON	1	8	7	0,93	court-circuit	Inconnue
2498		DAP	L34,5 kV	7	28	2,06	0,96	homopolaire	Défaut de terre à Bawku; défaut sur la ligne de
2499		LAF	Dép.CEET3	1	9	20,64	3,1	Surintensité	Pluie
2500	10/07/2015	MGL	TAG	4	301	20	100,33	court-circuit	Fluctuation de la tension TCN
2501		DJO	L1 SBEE	1	19	2,58	0,82	court circuit	Pluie
2502		PAR	T1	2	5	7,87	0,66	court-circuit	Défaut dans le réseau SBEE
2503		DAP	L34,5 kV	7	544	2,07	18,77	homopolaire	Défaut de terre à Bawku; temps pluvieux
2504		ATA	L34,5 kV	1	20	0,5	0,17	Surintensité	Pluie
2505		ATA	L440-KAR	1	4	16,44	1,1	court-circuit	Pluie
2506		CVE	T6	1	4	22	1,47	surintensité	Défaut dans le réseau SBEE
2507		AVA	Dép.OUIDAH	3	166	2,37	6,56	surintensité	déséquilibre
2508	16/07/2015	AVA	Dép.ALLADA	5	335	1,52	8,49	surcharge	déséquilibre

Figure 2 End of CEB operating data collection pages

We explored the data by reorganizing the faults by categories. Table 1 shows the types of faults and the causes that generated them. Also included in this table are the date, the location of the fault, the line on which the fault occurred, the number of trips, the power at interruption, the undistributed energy and the duration of interruption.

Table 1 Rearrangement of CEB operating data for characterization

Date	Name of Lines	Number of Sites	Trigger number	Power interrupted (MW)	Lost Energy Estimate (MWh)	Duration (mn)	Causes of failure	Type of defects
2010-09-13	CVE	L20-CGB	1	2	32	1,07	Overcurrent	Fault in the SBEE network
2013-03-26	MOM	L410-NAN	1	4	7,17	0,48	Short circuit	Voltage variation
2009-07-01	AVA	Dep. OUIDAH	1	5	2,5	0,21	Short circuit	Fault in the SBEE network
2012-08-16	SAK	T3	1	4	0,1	0,01	Short circuit	Disruption in the TCN network
2013-11-23	ONI	L3	6	26	0,75	0,33	Short circuit	Fault in the CEET network
2009-01-08	BOH	T1	1	18	2,5	0,75	Short circuit	Ground fault at Kara power plant
2013-03-19	DAP	L34,5	1	1	0,5	0,01	Short circuit	Sudden change in voltage
2008-07-30	ATA	L420-NAN	1	5	8,58	0,72	Overcurrent	Low oil level
2014-10-22	KARA	T1	2	8	6,5	0,87	Short circuit	Unknown
2013-07-25	LOK	Arrivée-L32	1	2	6,5	0,22	Overcurrent	Earth fault in Bawku
2008-11-25	MOM	L32-LOK	1	3	5,52	0,28	Overcurrent	Overload
2008-08-04	NAN	L410	1	60	4,25	4,25	Frequency fluctuation	Overload
2009-06-19	SAK	ATR2	1	0	36	4,25	Overcurrent	Fault in the SBEE network
2014-07-21	SAK	LA10	1	60	72,96	4,25	Overcurrent	Ground fault at Kara power plant
2008-09-21	AVA	Dép.OUIDAH	2	42	3	72,96	Short circuit	Fault in the SBEE network
2011-01-31	DAP	L34,5	7	479	0,6	2,1	Short circuit	Overload
2014-07-09	CVE	L20	1	4	36	4,79	Short circuit	Fault in the SBEE network
2008-08-10	LOK	Arrivée-L32	1	5	8,5	2,4	Overcurrent	Unknown
2010-04-27	KARA	T1	1	4	2	0,71	Overcurrent	Earth fault in Bawku
2012-05-14	SAK	TR3	4	24	0,1	0,13	Short circuit	Fault in the CEET network
2010-04-27	DAP	L34,5	7	42	3	0,04	Short circuit	Unknown

2010-09-13	CVE	L20	1	479	0,6	4,25	Overcurrent	Sudden change in voltage

To achieve this work, we first reorganized the failures based on the causes and we identified a total of 18. Then we used the linear kernel of Support Vector Machine, 100 iterations of logistic regression, 42 random states with 5 n_neighbors, 42 random states for random forest, 42 random states of gradient boosting and the ReLu activation function of artificial neural networks. Figure 3 presents the synopsis of the method and the organization of the steps. The models are subjected to training with 80% of the data and the remaining 20% were used to perform the tests following the Pareto law [23].

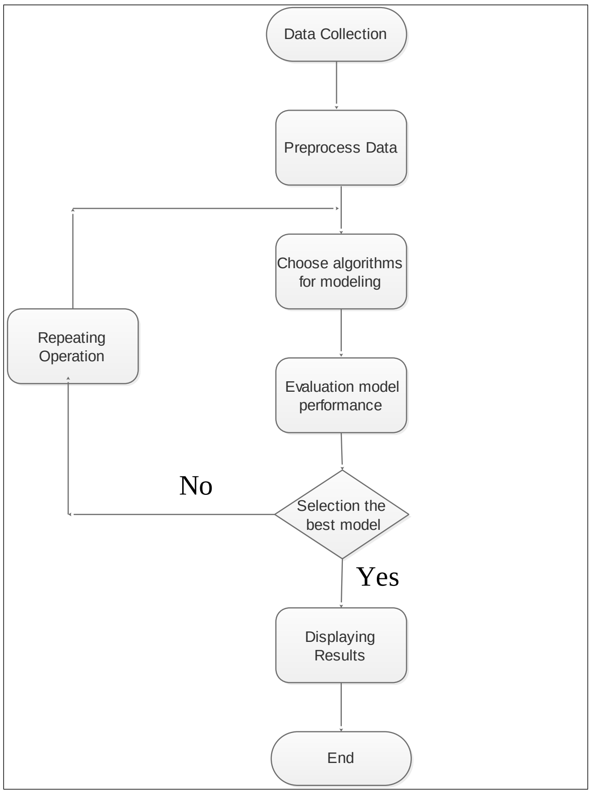


Figure 3 Synopsis and organization of the steps of the method

2.1. Support Vector Machine [24, 25]

Previous studies have employed machine learning for similar purposes. The approaches used included neural networks random forest and SVMs. SVMs were introduced by Vapnik and Chervonenkis (1981) and are widely used due to their flexibility in analyzing data with different distributions and their ability to deal with high-dimensional data such as gene expression. Previously SVMs using SNPs as predictors were employed for the classification of populations. In our implementation, we have chosen a linear kernel because it is suitable when the data are linearly separable. Indeed, it is sufficient to take a hyperplane that separates the classes, then to classify the data according to the side of the hyperplane where they are found. More formally, let a hyperplane separate the data. Then, it is sufficient to use the following function represented by the relation (1); sometimes called the indicator function; to perform the classification:

$$Classe(w.x+b) = Signe(w.x+b) \quad \text{..................(1)}$$

Where

$$Signe(w.x+b) = \begin{Bmatrix} -1, w.x+b<0 \\ 0, w.x+b=0 \\ 1, w.x+b>0 \end{Bmatrix}$$

2.2. Logistic regression [26, 27]

Logistic regression is a supervised classification model used to predict the probability that a sample belongs to a given class. It is particularly suited to binary classification problems, although it can be extended to multi-class cases via techniques such as multinomial logistic regression. In this work, logistic regression has been applied to predict failures in the high-voltage overhead transmission network. One of the main objectives of supervised learning is to provide a classification system that, for any new individual from the population, provides a prediction with accuracy if possible. Logistic regression can do this. But, unlike other methods, it can also provide an indicator of the reliability of the prediction with an estimate of the probability. Thus, when is close to 1 or 0, the prediction is rather safe; when it takes an intermediate value, close to the assignment threshold s (usually s = 0.5), the prediction is less safe. In areas where the consequences of misallocations can be dramatic, one could even imagine a system that only classifies with certainty by respecting the following conditions:

- if $\pi \leq s_1$ then $y = -$
- if $\pi \geq s_2$ then $y = +$
- else indeterminacy

2.3. K-Neirest_Neighbors [28, 29, 30, 31]

The K-Nearest Neighbors (KNN) model is a supervised learning method used for classification or regression tasks. The KNN algorithm works on the principle of proximity (see Figure 4). It classifies a new data sample according to the classes of the k closest data points in the training set. It has the advantage of simplicity, flexibility and does not make any assumptions about the data distribution. On the other hand, it has limitations such as: high computational cost; intensive memory; sensitivity to dimensions and sensitivity to noisy data.

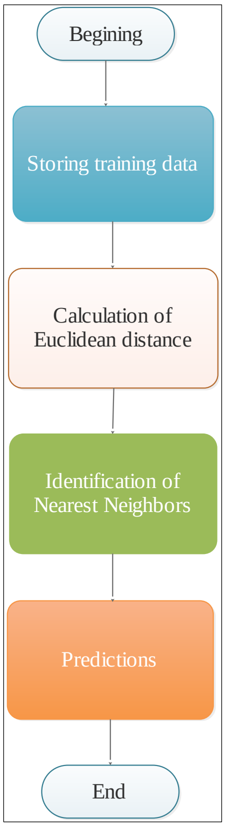

Begining
Storing training data
Calculation of Euclidean distance
Identification of Nearest Neighbors
Predictions
End

Figure 4 K-Nearest Neighbors Operational Flowchart

2.4. Random forest, [32, 33]

The Random Forest Classifier model is an ensemble-based machine learning algorithm that uses multiple decision trees to improve prediction accuracy and reduce the risk of overfitting. First, the model is initialized with a random_state parameter. This ensures reproducibility of the results. The model is then trained on the training set. A random forest consists of aggregating the prediction of several trees. The idea behind this technique is to group the mean (in the case of regression) of the predictions in order to reduce the variance associated with it. The principle consists of aggregating the prediction of several different regression trees [17]. Figure 5 shows how its algorithm works.

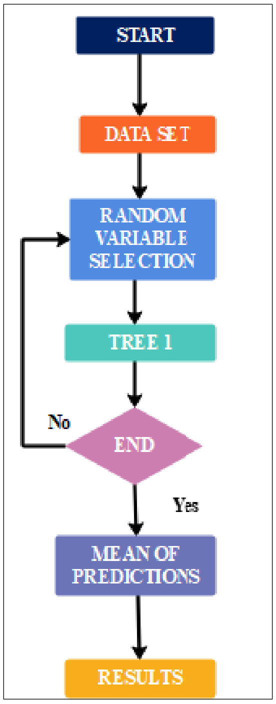


Figure 5 Flowchart of Random Forest operation

2.5. Gradient Boosting, [34, 35, 36, 37, 38]

The Gradient Boosting Classifier model is an ensemble method that builds decision trees iteratively. Each new tree corrects the errors of the previous trees. This improves the overall performance of the model. To optimize the performance of the model, a grid search is performed. This allows exploring different combinations of hyper parameters. The hyper parameters tested include:

- n_estimators: Number of trees in the model (tested values: 50, 100);
- max_depth: Maximum depth of trees (tested values: 3, 5);
- min_samples_split: Minimum number of samples needed to split a node (tested values: 5, 10);
- min_samples_leaf: Minimum number of samples needed to be a leaf (tested values: 2, 5);
- learning_rate: Learning rate to control the contribution of each tree (tested values: 0.01, 0.1);
- subsample: Proportion of samples to use for training (tested value: 0.8).

2.6. ReLu Function of Artificial Neural Networks, [39, 40, 41]

Artificial neural networks are highly connected networks of elementary processors operating in parallel (neurons), each artificial neuron is an elementary processor, it receives a variable number of inputs from upstream neurons, each of these inputs is associated with a weight W representing the connection strength, each elementary processor calculates a single output based on the information it receives, which then branches out to feed a variable number of downstream neurons. Each connection is associated with a weight. It is possible to improve the efficiency of the processing by inserting between the processing layers a layer that will operate a mathematical function (activation function) on the output signals. The ReLU (Rectified Linear Unit) function is formulated by the relation (2).

$$Y(x) = \max(0, x) \qquad .... (2)$$

This function forces neurons to return positive values. Artificial Neural Network (ANN) is a machine learning model inspired by the functioning of the human brain. In this work, we designed a neural network model to predict faults in the high-voltage overhead transmission network. The model is built according to a sequential architecture and consists of the following elements: an Input Layer, 99 Hidden Layers and an Output Layer.

2.7. Performance evaluation criteria [42]

In the context of the binary classification problem, there are four possible cases that occur: a true positive prediction (TP), a true negative prediction (TN), a false positive prediction (FP), and a false negative prediction (FN). Based on the number of times these four cases occur, that make up the confusion matrix, many different performance measures have been proposed. Amongst these measures there are accuracy (ACC), precision (PPV), recall (TPR), false negative rate (FNR), false positive rate (FPR), specificity (TNR), prevalence (PT) and F1 score (F1); their definitions are listed in Table 2.

Table 2 Summary of metrics and their determination formula

Metrics	**Definitions**	**Determination formulas**
ACC	Accuracy	$ACC = \frac{TP + TN}{TP + TN + FP + FN}$
PPV	Precision	$PPV = \frac{TP}{TP + FP}$
TPR	Recall	$TPR = \frac{TP}{TP + FN}$
FNR	False negative rate	$FPR = \frac{FP}{FP + TN}$
FPR	False positive rate	$FPR = \frac{FP}{FP + TN}$
TNR	Specificity	$TNR = \frac{TN}{TN + FP}$
F1	F1 Score	$F1 = \frac{2TP}{2TP + FP + FN}$

For a binary classification, the confusion matrix has the theoretical form observed in Figure 6. The confusion matrix is a tool that measures the performance of a classification model with two or more classes, which allows the observed values to be compared with those of the prediction and is used to verify the correct classification of the data.

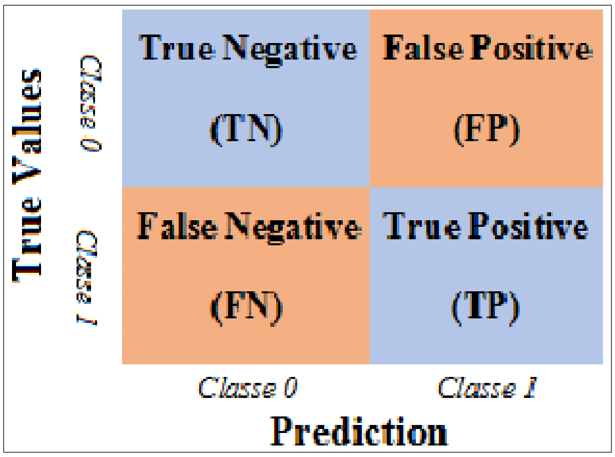


Figure 6 Graphical view of the confusion matrix sections

The accuracy measure gives the user an overall view of the model's performance, all classes combined. This involves establishing the proportion of correctly classified examples, all classes combined, among all instances. The precision measure, on the other hand, makes it possible to evaluate to what extent the model is correct in its predictions, regardless of the class. Concretely, this measure gives the proportion of correctly classified examples, all classes combined. The recall, on the other hand, corresponds to the proportion of positive examples correctly predicted by the model. The higher it is, i.e. close to 1, the more the model is able to classify the positives. In addition, we will also plot the ROC curve to refine the results. This is a tool for evaluating and comparing models. The area under the curve (AUC) is a parameterized function of sensitivity and specificity as a function of the threshold varying between 0 and 1. The ROC curve is therefore plotted using two variables: a binary variable and a continuous one, [43]. The ROC curve has many advantages among which we can list the following:

- it is independent of cost and misallocation matrices and allows to know if one model will always be better than the other; whatever the cost matrix;
- it is operational even in the case of very unbalanced distributions;
- its results remain valid even if the test sample is not representative

3. Results

Following the classification, 19 causes of failures were noted. Among these, short circuit takes the lead with 39.019% or 947 occurrences out of 2427 during the study period. It is followed by overcurrent which amounts to 889 corresponding to 36.629%. On the other hand, loss of excitation, voltage drop and burning of the disconnector blade remain very rare (0.123% corresponding to 3 occurrences for 2427 encountered during the study period). The results of the classification are cumulated in Table 3.

Table 3 Classification of causes of failures by number of occurrences

Cause of breakdowns	Number of appearances	Frequency (%)
Short circuit	947	39,0193655
Overcurrent	889	36,6295838
Differential	123	5,06798517
None	97	3,99670375
Homopolar	95	3,91429749
Network collapse	92	3,79068809
Overload	83	3,41985991
Central trigger	22	0,90646889
Loss of voltage	20	0,82406263
Overvoltage and undervoltage	13	0,53564071
Meter explosion	9	0,37082818
Mini frequency	9	0,37082818
Max frequency	8	0,32962505
Imbalance	7	0,28842192
Lack of tension	4	0,16481253
Cutter knife burn	3	0,12360939
Voltage drop	3	0,12360939
Loss of excitement	3	0,12360939
Total	2427	100

Concerning the modeling, the confusion matrices are well seen through the figures ranging from 7 to 12. Figure 7 represents the results of K-Nearest Neighbors; 8 that of logistic regression; 9 gives for the Support Vector Machine. In Figure 10, it is about Random Forest; 11 presents for Gradient boosting and finally the Figure 12 represents the confusion matrix of the modeling by artificial neural networks. Table 4 summarizes the values obtained by class and by algorithms.

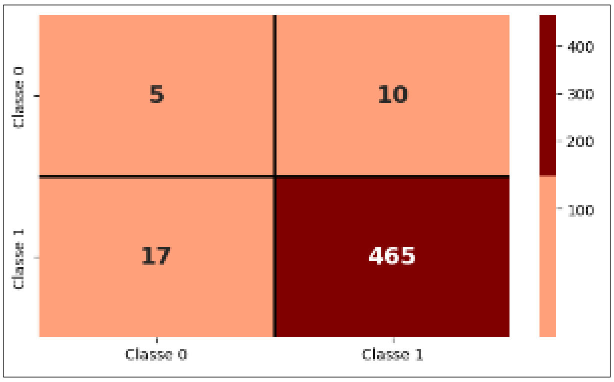


Figure 7 Confusion matrix of K-Nearest Neighbors

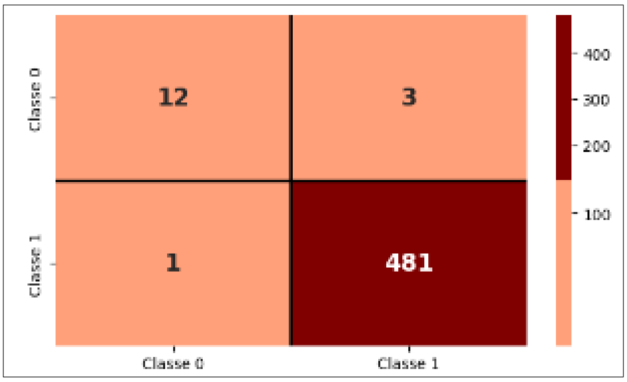


Figure 8 Confusion matrix of Logistic Regression

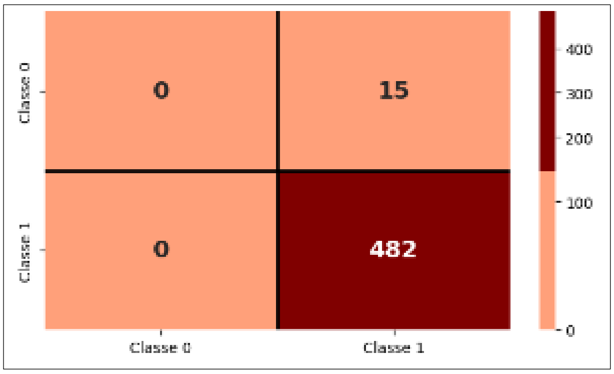


Figure 9 Confusion matrix of Support Vector Machine

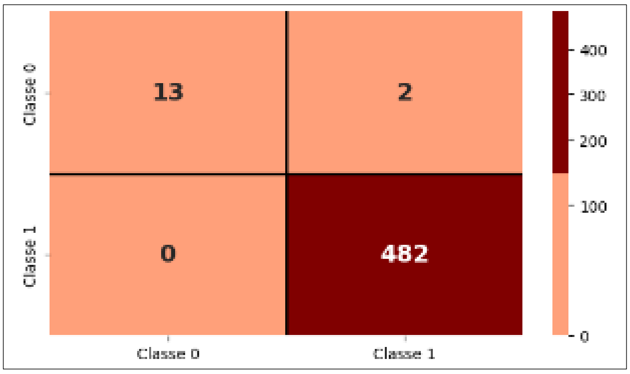


Figure 10 Confusion matrix of Random Forest

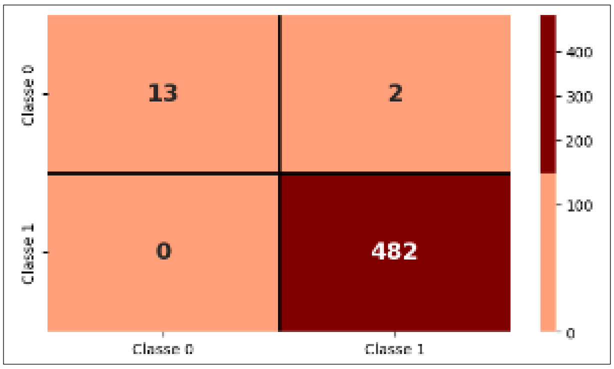


Figure 11 Confusion matrix of Gradient Boosting

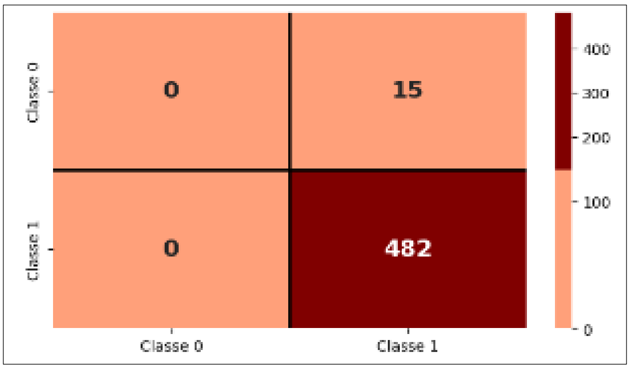


Figure 12 Confusion matrix of Réseaux de Neurone Artificiels

Table 4 Cumulative values of the confusion matrix by class and by algorithm

Modeling Algorithms	Cumulative values by class			
	False Positive	False Negative	True Negative	True Positive
K-Nearest Neighbors (KNN)	10	17	5	465
Logistic Regression	3	1	12	451
Support Vector Machine	15	0	0	482
Random Forest	2	0	13	482
Gradient Boosting Classifier	2	0	13	482
Artificial Neural Network (ANN)	15	0	0	482

Regarding the ROC curves, the graphs in Figure 13 show the performances. In Table 5, we find the numerical results of the Scores and the diagrams in Figure 14 give the graphical evaluation of the performances by explored model.

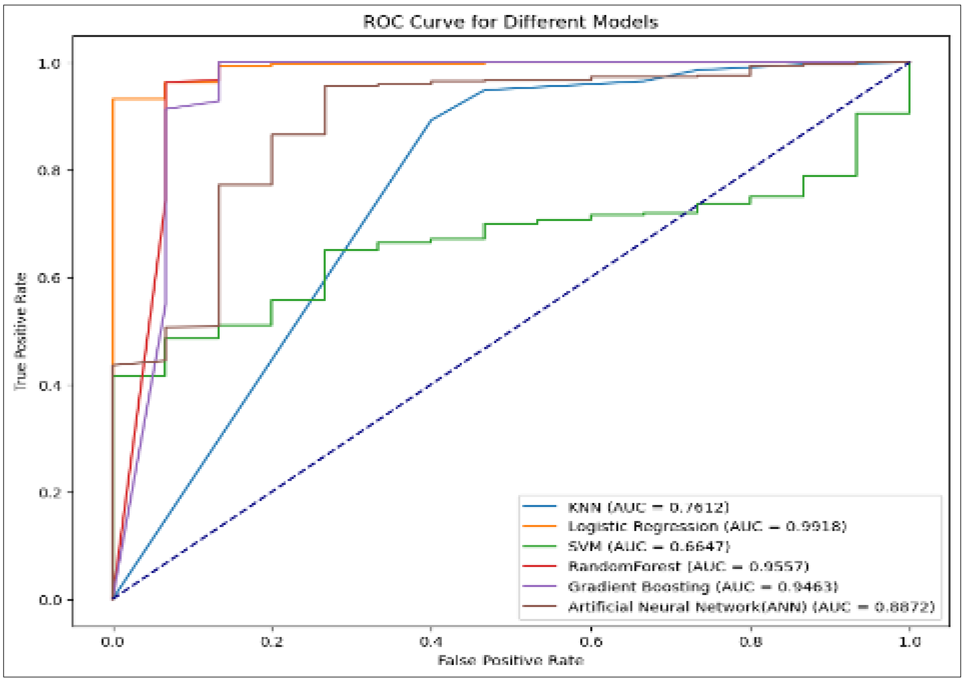


Figure 13 Graphical view of the ROC Curve of the explored Models

Table 5 Results of the evaluation criteria and the ROC curve of the models by algorithm

Modeling Algorithms	Results of the Model Evaluation Criteria				ROC AUC
	Accuracy	F1 Score	Precision	Recall	
Logistic Regression	0.991952	0.991673	0.991667	0.991952	0.9918
SVM	0.989262	0.988552	0.988916	0.989262	0.6647
Nearest Neighbors (KNN)	0.955734	0.957072	0.958491	0.955734	0.7612
Random Forest	0.991946	0.991414	0.992013	0.991946	0.9557
Gradient Boosting Classifier	0.995973	0.995933	0.995921	0.995973	0.9463
Artificial Neural Network (ANN)	0.966443	0.949951	0.934012	0.966443	0.8872

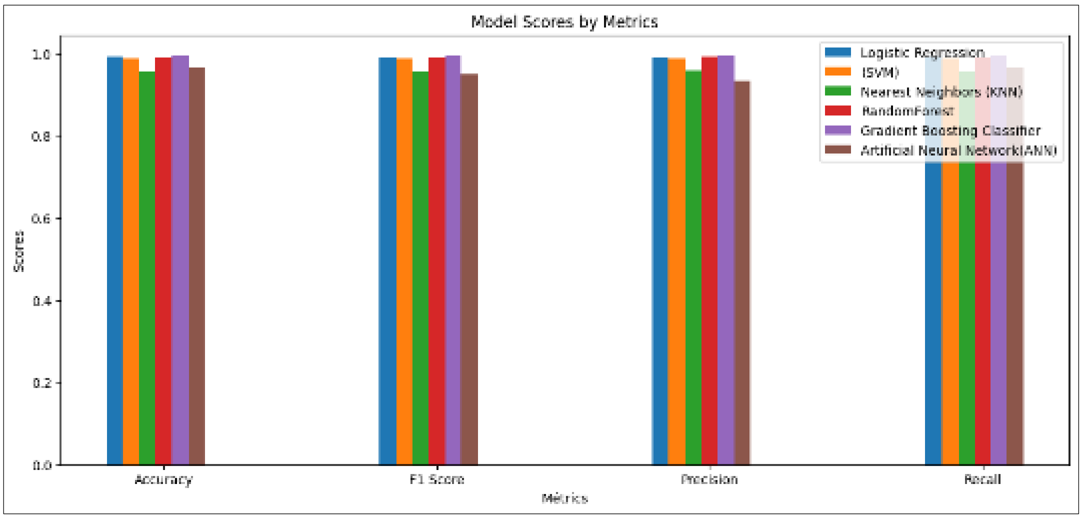


Figure 14 Bar chart of the different models

Table 6 Cumulative results by distribution of confusion matrix values, ROC curve and performance evaluation metrics

Algorithms	Distribution by class				Model Evaluation Results				ROC AUC
	FP	FN	TN	TP	Accuracy	F1 Score	Precision	Recall	
Logistic Regression	3	1	12	451	0.991	0.991	0.991	0.991	0.991
SVM	15	0	0	482	0.989	0.988	0.988	0.989	0.664
Nearest Neighbors (KNN)	10	17	5	465	0.955	0.957	0.958	0.955	0.761
Random Forest	2	0	13	482	0.991	0.991	0.992	0.991	0.955
Gradient Boosting Classifier	2	0	13	482	0.995	0.995	0.995	0.995	0.946
Artificial Neural Network (ANN)	15	0	0	482	0.966	0.949	0.934	0.966	0.887

4. Discussion

An abnormal increase or decrease in nominal values in an electrical circuit constitutes a fault or disturbance. The main objective of protection is to eliminate the fault. Depending on the type of fault, intelligent protection emits a circuit breaker trip signal and consequently the powering down of the installation or a signaling signal to inform operators of the nature of the fault allowing them to take exact measurements. Intended to prevent equipment from being passed through by currents harmful to itself and its environment, protection devices must cut off the circuit under load in such circumstances. The implementation of protection of electrical installations requires different equipment, operations and actions, with specific functions, which must be perfectly mastered. Their nature depends on: the type of protection targeted (protection against overloads, protection against short circuits, etc.) and their capacity to ensure this protection, [44, 45].

It is important to note that despite the effectiveness of these protection devices, we still encounter all kinds of faults that cause power interruptions. As shown in Table 1 of this document for the case of the CEB network. Given that the objective of this work is to predict faults in order to take steps to avoid their occurrence, we have carried out a characterization in Table 3 and the analysis shows us that there are so far 19 different faults in the network studied that cause power interruptions. The predictive modeling of these faults, carried out with some algorithms of supervised learning of artificial intelligence allows us to confirm that we can be warned in order to carry out operations and avoid interruptions.

At first glance, if we refer to the scores, all the models present very interesting results. The artificial neural networks which gives the lowest is 0.949. Then comes the K-Nearest Neighbors which amounts to 0.957, confirmed by the SVM whose value is 0.988. The remainders (Gradient Boosting, Random Forest and Logistic Regressions) give scores higher than 0.99. Indeed, most classifier models do not produce only a binary classification as a result. They generally calculate a score to classify cases as positive or negative. The score is generally converted into a percentage. This score does not always necessarily imply a true probability, especially for machine learning techniques, but often indicates a ranking of cases, rather than a strict probability. Since each case would have a score, it would be appropriate to assign a threshold beyond which the model result could be considered positive. By the way, a case is considered standard only when the score is greater than 80%. We obtained them for all the models studied. Moreover, if we have two models predicting with standard cases or not, we can tabulate a confusion matrix for each model. This being done, the confusion matrices obtained through the figures ranging from 7 to 12, allowed us to validate the predictions.

From our confusion matrices, we can estimate that there were 497 cases during training and distributed across classes. Our sensitivity is then 100% and our accuracy seems satisfactory at 93.56% for Nearest Neighbors; 96.78% for logistic regression then varying between 80% and 96.98% for the other models. We then deduce that our model predicts that all cases will be standard.

Furthermore, the recall allows us to know the percentage of True Positives predicted by the models. In other words, it is the number of predicted True Positives divided by all the positives (True Positive + False Negative). The interest of the recall lies in the fact that the higher it is, the more the Machine Learning model maximizes the number of True Positives. This observation is also made in the results with 0.966 for Artificial Neural Networks and for SVMs, its value is 0.989. The other algorithms give recall values greater than 0.99. It should be noted, however, that when the recall is high, it rather means that it will not miss any positives. However, this does not provide any information on its prediction quality on negatives. To confront this remark, the ROC curves provide more precision. All the curves are found at the upper part of the reference axis $y = x$.

The Receiver Operating Characteristic (ROC) curve plots the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) as the decision threshold varies. The ROC curve can be less useful when training models on datasets with high class imbalance, as the majority class can be swamped by minority classes. In reality, the area under the curve can be interpreted as the proportion of samples correctly classified. More precisely, it represents the probability that the classifier will classify a randomly selected positive sample at a higher rank than a randomly selected negative sample. The shape of the curve gives an indication of the relationship between the True Positive Rate and False Positive Rate values as a function of the classification threshold or decision boundary. We observe in Figure 13 that all the curves approach the upper left corner of the graph. Which leads to True Positive Rate values that tend towards 100% and False Positive Rate values of 0%. Which corresponds to the best possible model because a random model produces a ROC curve along the line $y = x$ from the lower left corner to the upper right corner. A model worse than random would have a ROC curve that passes under the line and this is what the case of the model obtained with SVM modeling presents. However, it is a very small portion that presents itself this way, revaluing the SVM forecast like the other algorithms explored and confirmed by the histograms in Figure 14 which are almost at the same level. It goes without saying that the results obtained through this work are satisfactory, thus giving the Electricity Community of Benin the opportunity to move on to its implementation for future decisions

5. Conclusion

They will allow us to carry out the learning of the failures noted in the network of the Electric Community of Benin (CEB). The objective is to predict the failures in order to prevent the faults from creating interruptions in the supply of electricity. As materials we have the operating data and the recurrence of the faults in the network from 2008 to 2015. As methods, we used algorithms such as SVM, K-Nearest Neighbors, Random Forest, Gradient Boosting, Artificial Neural Network and Logistic Regression. The results are subjected to evaluation criteria such as the confusion matrix, the ROC curve and the scores (Accuracy, F1 Score, Precision and recall). A characterization of the failures is carried out.

The results of the characterization show that there are 19 different faults that cause interruptions in the studied network. Among these, the most recurrent in descending order are: short circuit, overcurrent and differential. It was noted that the cause of the faults caused by the short circuit amounts to 947 out of 2427, corresponding to a frequency of 39.019%. There are 889 overcurrent worth 36.629% and 123 out of 2427 which is equivalent to 5.067% for the differential that also created the interruption of the current in the network. In contrast to these recurring causes, we find: burning of the disconnector knife; voltage drop and loss of excitation; which only appeared 3 times per case out of 2427, during the study period on the network worth together 0.369%.

Concerning the modeling, all the results are very interesting. Only, some are better than others. The most unfavorable result obtained through the area under the ROC curve, noted AUC is 0.761 for the K-Nearest Neighbors. Its score (Accuracy = 0.955; F1 Score = 0.957; Precision = 0.958 and Recall = 0.955) confirms this observation. The area under the ROC curve (AUC), remains even more unfavorable with the modeling by Support Vector Machine (0.664) but its score (Accuracy = 0.989; F1 Score = 0.988; Precision = 0.988 and Recall = 0.989) exceeds that of K-Nearest Neighbors. Furthermore, for the other models, the AUC exceeds 80% with the logistic regression at the head giving AUC = 0.991 and (Accuracy = 0.991; F1 Score = 0.991; Precision = 0.991 and Recall = 0.991) as score. Indeed, the ROC (Receiver Operating Characteristics) curve [Fawcett (2003)] offers both a graphical vision and a relevant measure of the performance of a classifier. It has many advantages over recall and precision measures by class: the performance is synthesized by a single measure that does not depend on the class proportions. Recall and precision measures are also useful because they precisely characterize the behavior of the classifier on each of the classes.

The values of True Positives that remain greater than 450 out of 497, for the confusion matrices by classes, confirm these results across all the algorithms explored. It is 451 for logistic regression, 465 for K-Nearest Neighbors and 482 for the rest of the algorithms. Given that the False Positives have gradually become True Negatives without affecting the true positives, we can conclude that the models are perfect. This demonstrates the validity of the approaches, thus opening the door to forecasts of maintenance activities; while theoretically reducing electricity interruptions in the CEB network.

However, it is necessary to experiment in the future work, the results over the periods recognized that optimizations and adjustments are possible to further improve the proposed solution. It is also important to resume the work taking into account the new data since the network is expanding at every moment.

Compliance with ethical standards

Acknowledgments

Our thanks go to the Benin Electric Community (CEB) for its availability, its collaboration and its confidence in our work by entrusting it with this study and by providing it with the data necessary for the completion of this research work.

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Bernard Tissot, Primary energy sources and the greenhouse effect, Geoscience Proceedings, 2003, Volume 335, Issues 6–7, ISSN 1631-0713, Pages 597-601, https://doi.org/10.1016/S1631-0713(03)00104-4.

[2] Claude Seyer The evolution of consumption and production of different energy sources in France between 1970 and 1980, Eastern Geographical Review, 1980, 20-1-2, pp. 43-64

Balat, M., Ayar, G., Oguzhan, C., Uluduz, H., & Faiz, U.. Influence of Fossil Energy Applications on Environmental Pollution. Energy Sources, Part B: Economics, Planning, and Policy, 2007, 2(3), 213–226. https://doi.org/10.1080/15567240500402768

[3] Mustafa Tevfik Kartal, The role of consumption of energy, fossil sources, nuclear energy, and renewable energy on environmental degradation in top-five carbon producing countries, Renewable Energy, 2022, Volume 184, Pages 871-880, ISSN 0960-1481, https://doi.org/10.1016/j.renene.2021.12.022.

[4] Michael A. Mac Kinnon, Jacob Brouwer, Scott Samuelsen, The role of natural gas and its infrastructure in mitigating greenhouse gas emissions, improving regional air quality, and renewable resource integration, Progress in Energy and Combustion Science, 2018, Volume 64, Pages 62-92, ISSN 0360-1285, https://doi.org/10.1016/j.pecs.2017.10.002.

[5] Ottinger, Richard L., and Rebecca Williams. "RENEWABLE ENERGY SOURCES FOR DEVELOPMENT." Environmental Law, 3 Dec. 2024, vol. 32, no. 2, 2002, pp. 331–68, JSTOR, http://www.jstor.org/stable/43267559.

[6] Ibrahim Yuksel, Kamil Kaygusuz, Renewable energy sources for clean and sustainable energy policies in Turkey, Renewable and Sustainable Energy Reviews, 2011, Volume 15, Issue 8, Pages 4132-4144, ISSN 1364-0321, https://doi.org/10.1016/j.rser.2011.07.007.

[7] M. Moazzami, R. Hemmati, F. Haghighatdar Fesharaki, S. Rafiee Rad, Reliability evaluation for different power plant busbar layouts by using sequential Monte Carlo simulation, International Journal of Electrical Power & Energy Systems, 2013, Volume 53, Pages 987-993, ISSN 0142-0615, https://doi.org/10.1016/j.ijepes.2013.06.019.

[8] Hatice Yılmaz, Characterization and comparison of leaching behaviors of fly ash samples from three different power plants in Turkey, Fuel Processing Technology, 2015, Volume 137, Pages 240-249, ISSN 0378-3820, https://doi.org/10.1016/j.fuproc.2015.04.011.

[9] Edouard González-Roubaud, David Pérez-Osorio, Cristina Prieto, Review of commercial thermal energy storage in concentrated solar power plants: Steam vs. molten salts, Renewable and Sustainable Energy Reviews, 2017, Volume 80, Pages 133-148, ISSN 1364-0321, https://doi.org/10.1016/j.rser.2017.05.084.

[10] S.N. Singh : Electric power generation, transmission and distribution, second edition, PHI Learning, Private limited, New delhi-110001, ISBN-978-81-203-3560-8, 2011

[11] R. Strzelecki and G. Benysek : Power electronics in smart electronical energy network, power systems-series, springer-vertag London Limited, eISBN-978-1-84800-318-7, 2008

[12] Rahman Dashti, Mohammad Daisy, Hamid Mirshekali, Hamid Reza Shaker, Mahmood Hosseini Aliabadi, A survey of fault prediction and location methods in electrical energy distribution networks, Measurement, 2021, Volume 184, 109947, ISSN 0263-2241, https://doi.org/10.1016/j.measurement.2021.109947.

[13] Maduako, I., Igwe, C.F., Abah, J.E. et al. Deep learning for component fault detection in electricity transmission lines. J Big Data 9, 81, 2022. https://doi.org/10.1186/s40537-022-00630-2

[14] S. Zhang, Y. Wang, M. Liu and Z. Bao, "Data-Based Line Trip Fault Prediction in Power Systems Using LSTM Networks and SVM," in IEEE Access, 2018, vol. 6, pp. 7675-7686, doi: 10.1109/ACCESS.2017.2785763.

[15] Renga, D., Apiletti, D., Giordano, D. et al. Data-driven exploratory models of an electric distribution network for fault prediction and diagnosis. Computing 102, 2020, 1199–1211, https://doi.org/10.1007/s00607-019-00781-w

[16] Berry, Michael W., Azlinah Mohamed, and Bee Wah Yap, eds. Supervised and unsupervised learning for data science. Springer Nature, 2019.

[17] SHINDE, Pramila P. et SHAH, Seema. A review of machine learning and deep learning applications. In : 2018 Fourth international conference on computing communication control and automation (ICCUBEA). IEEE, 2018, p. 1-6.

[18] Oloulade, A., Imano, A. M., Vianou, A., & Badarou, R.. Contribution to the study of power distribution and the evaluation of losses in the transport and distribution networks of the Benin electrical community and the Beninese electric energy company (CEB-SBEE). Sciences, Technologies and Development (July 2016).

[19] GUENOUPKATI, Agbassou, SALAMI, Adekunlé Akim, KODJO, Mawugno Koffi, et al. Short-Term Electricity Load Forecasting Using K-Means Clustering-Artificial Neural Networks Hybrid Model: Case Study Of Benin Electricity Community (CEB). In : 2021 IV International Conference on High Technology for Sustainable Development (HiTech). IEEE, p. 01-05.

[20] Salami, A. A., Ajavon, A. A., Dotche, K. A., & Bedja, K. S.. Electrical load forecasting using artificial neural network: The case study of the grid inter-connected network of benin electricity community (CEB). Am. J. Eng. Appl. Sci, 2018,11(2), 471-481.

[21] NTAGUNGIRA, Carpophore. Problems of access to electricity in Togo. 2015.

[22] Hardy M. : Pareto's law. The Mathematical Intelligencer, 32, 38-43, 2010.

[23] Timothy Vivian-Griffiths, and al. "Predictive modeling of schizophrenia from genomic data: Comparison of polygenic risk score with kernel support vector machines approach" American Journal of Medical Genetics, Part B, Neuropsychiatric Genetics, 04 December 2018, https://doi.org/10.1002/ajmg.b.32705

[24] Ashanira Mat Deris, Azlan Mohd Zain, Roselina Sallehuddin, " Overview of Support Vector Machine in Modeling Machining Performances", Procedia Engineering, 2011, Volume 24, Pages 308-312, ISSN 1877-7058, https://doi.org/10.1016/j.proeng.2011.11.2647.

[25] Adeline Gillet, Yves Brostaux, Rodolphe Palm "Main models used in logistic regression", Biotechnol. Agron. Soc. Environ. 2011 15 (3), 425-433, https://hdl.handle.net/2268/112603

[26] CHIABRI, L., ICHOU R., BENTAHAR A., “Evaluation of the impact of the territorialization of the Green Morocco Plan on the standard of living of phoeniculists in the oasis areas of the Drâa-Tafilalet region through a binary logistic regression” Economic Managerial Alternatives, 2023, Vol 5, Special issue 2 168-185, E-ISSN: 2665-7511, June, https://revues.imist.ma/?journal=AME

[27] ZHANG, S., LI, X., ZONG, M., ZHU, X., WANG, R., « Efficient kNN Classification With Different Numbers of Nearest Neighbors », IEEE Transactions on Neural Networks and Learning Systems, Vol. 29, No. 5, pp. 1774-1785, May 2018 https://ieeexplore.ieee.org/document/7927711

[28] Ni Li ,Haipeng-Kong ,Yaofei Ma , Gong de Guanghong et Wenqing Huai « Human Performance Modeling for Manufacturing Based on Improved KNN Algorithm”, International Journal of Advanced Manufacturing Technologies, 2016, Volume 84, pages 473–483,https://link.springer.com/article/10.1007/s00170-016-8418-6

[29] Zhang Shichao « Cost-sensitive KNN classification », Neuroinformatique, 28 mai 2020, Volume 391, Pages 234-242, https://doi.org/10.1016/j.neucom.2018.11.101

[30] A. Moldagulova and R. B. Sulaiman, "Using KNN algorithm for classification of textual documents," 2017 8th International Conference on Information Technology (ICIT), Amman, Jordan, 2017, pp. 665-671, doi: 10.1109/ICITECH.2017.8079924.

[31] Torgyn Shaikhina, Dave Lowe, Sunil Daga, David Briggs, Robert Higgins, Natasha Khovanova « Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation »Biomedical Signal Processing and Control, 2019, Volume 52, Pages 456-462, ISSN 1746-8094, https://doi.org/10.1016/j.bspc.2017.01.012

[32] Komla Kpomonè Apaloo Bara, Eyouleki Tcheyi Gnadi Palanga, Komi Ghislain Ekegnon and Koffi-Sa Bedja : Naive Bayesian classifier and random forest approaches for modeling the electrical resistivity of soils in tropical zones by meteorological variables: case of nine sites in Lomé, Togo, World Journal of Advanced Research and Reviews, 2023, 20(03), 037–050,

[33] Adeniyi J. Adewale,Irina Dinu &Yutaka Yasui « Boosting for correlated binary classification », Journal de statistique computationnelle et graphique Volume 19, - Numéro 1, 2010. https://doi.org/10.1198/jcgs.2009.07118

[34] Bahad, P., Saxena « Étude des algorithmes AdaBoost et Gradient Boosting pour l'analyse prédictive » Dans : Singh Tomar, G., Chaudhari, NS, Barbosa, JLV, Aghwariya, MK (éd.) Conférence internationale sur l'informatique intelligente et la communication intelligente, 2019. https://doi.org/10.1007/978-981-15-0633-8_22

[35] Leo Guelman « Gradient boosting trees for auto insurance loss cost modeling and prediction », Expert Systems with Applications, 2012. Volume 39, Issue 3, Pages 3659-3667, ISSN 0957-4174,

[36] https://doi.org/10.1016/j.eswa.2011.09.058

[37] Juan Pineda-Jaramillo, Ph.D. and Óscar Arbeláez-Arenas « Assessing the Performance of Gradient-Boosting Models for Predicting the Travel Mode Choice Using Household Survey Data », Journal of Urban Planning and Development, June 2022 Volume 148, Issue 2 . https://doi.org/10.1061/(ASCE)UP.1943-5444.0000830

[38] Yung-Chia Chang, Kuei-Hu Chang, Guan-Jhih Wu « Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions », Applied Soft Computing, 2018Volume 73, Pages 914-920, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2018.09.029

[39] Ricardo de A. Araújo, Adriano L.I. Oliveira, Silvio Meira « A morphological neural network for binary classification problems »,Engineering Applications of Artificial Intelligence,, 2017, Volume 65, Pages 12-28, ISSN 0952-1976 https://doi.org/10.1016/j.engappai.2017.07.014

[40] Lkhagvadorj Munkhdalai, Tsendsuren Munkhdalai, Keun Ho Ryu, « GEV-NN: A deep neural network architecture for class imbalance problem in binary classification », Knowledge-Based Systems, 2020, Volume 194, 105534, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2020.105534

[41] Y. Alparslan et al., « Towards Searching Efficient and Accurate Neural Network Architectures in Binary Classification Problems, », International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, pp. 1-8, 2021. DOI: 10.1109/IJCNN52387.2021.9533483.

[42] Kevin Riehl, Michael Neunteufel, Martin Hemberg : Hierarchical confusion matrix for classification performance evaluation, June 2023, DOI: 10.48550/arXiv.2306.09461

[43] Loic Desquilbet. Tutorial on ROC curves and their creation using the easyROC website. hal-02870055v2, 2022. https://hal.science/hal-02870055v2

[44] Mircea, P.M., Electrical Safety in LV Energy Installations. In: Lazaroiu, G.C., Roscia, M., Dancu, V.S. (eds) Energy Transition Holistic Impact Challenge (ETHIC): A New Environmental and Climatic Era. Environmental Science and Engineering. Springer, Cham, (2024). https://doi.org/10.1007/978-3-031-55448-3_16

[45] G. Narasimha, "LV protective devices for electrical equipment," Conference Record of the 2002 IEEE Industry Applications Conference. 37th IAS Annual Meeting (Cat. No.02CH37344), Pittsburgh, PA, USA, 2002, pp. 2252-2257 vol.3, doi: 10.1109/IAS.2002.1043846.

[46] François Rioult. "Graphical interpretation of the ROC curve. Knowledge Extraction and Management" (EGC’11), Jan 2011, Brest, France. 6 p. hal-01018456