APLICAÇÃO DE *MACHINE LEARNING* PARA CLASSIFICAÇÃO DE DESCARGAS PARCIAIS EM REDES DE DISTRIBUIÇÃO DE ENERGIA ELÉTRICA SUBTERRÂNEA

APPLICATION OF MACHINE LEARNING TO CLASSIFY PARTIAL DISCHARGES IN UNDERGROUND ELECTRICAL ENERGY DISTRIBUTION NETWORKS

Eliane Suely Everling Paim

Instituto Federal Catarinense, Concórdia, SC, Brasil
Mestra em Modelagem Matemática. E-mail: eliane.paim@sou.unijui.edu.br
https://orcid.org/0000-0002-9775-1753

Maurício de Campos

Universidade do Vale do Itajaí, Itajaí, SC, Brasil
Doutor em Engenharia Elétrica. E-mail: campos@unijui.edu.br
https://orcid.org/0000-0001-6499-4145

Airam Teresa Zago Romcy Sausen

Universidade Regional do Noroeste do Estado do Rio Grande do Sul, Ijuí, RS, Brasil
Doutora em Engenharia Elétrica. E-mail: airam@unijui.edu.br
https://orcid.org/0000-0001-6499-4145

Paulo Sérgio Sausen

Universidade Regional do Noroeste do Estado do Rio Grande do Sul, Ijuí, RS, Brasil
Doutor em Engenharia Elétrica. E-mail: sausen@unijui.edu.br
https://orcid.org/0000-0001-9863-8800

Submissão: 23-11-2023 Aceite: 11-06-2025

Resumo: A deterioração devido ao envelhecimento e às descargas parciais está entre as principais causas das faltas em cabos de energia elétrica. As técnicas de *Machine Learning* despontam como um diferencial no reconhecimento automatizado de descargas parciais. Portanto, neste trabalho são investigados e comparados vários métodos de *machine learning* como estratégia para classificar e detectar as descargas parciais em sistemas de distribuição de energia elétrica subterrânea. A partir do modelo de um dos padrões de descarga parcial de uma rede real é investigada a capacidade de classificação das descargas parciais em redes de distribuição de energia subterrânea, utilizando *machine learning*. Foram elaboradas simulações utilizando o software Matlab®. As taxas de acurácia, sensibilidade e especificidade foram avaliadas e comparadas para cinco modelos de classificadores, assim como taxas

Revista Vivências | Erechim | v. 21 | n. 43 | p. 61-77 | jul./agos. 2025
DOI: https://doi.org/10.31512/vivencias.v21i43.1242

de erro. Os resultados mostraram-se satisfatórios para classificação das descargas parciais quando comparados com resultados da literatura que aborda problema correlato. Ainda que este trabalho se encontre em fase inicial, com resultados parciais, esses resultados já apontam para uma estratégia viável.

Palavras-chave: *Machine learning*. Classificadores. Descargas parciais. Redes subterrâneas.

Abstract: Deterioration due to aging and partial discharges are the main causes of faults in electrical power cables. Machine learning techniques emerge as a differentiator in the automated recognition of partial discharges. Therefore, in this work, several Machine learning methods are investigated and compared as a strategy to classify and detect partial discharges in underground electrical energy distribution systems. Using the model of one of the partial discharge patterns of a real network, the ability to classify partial discharges in underground energy distribution networks is investigated, using Machine learning. Simulations were created using Matlab® software. Accuracy, sensitivity and specificity rates were evaluated and compared for five classifier models, as well as error rates. The results were satisfactory for classifying partial discharges, when compared with results from the literature that addresses a related problem. Although this work is in its initial phase with partial results, they already point to a viable strategy.

Keywords: Machine Learning. Classifiers. Partial discharges. Underground networks.

Introdução

Percebe-se, atualmente, em sistemas de distribuição de alta e média tensão, um incremento da utilização de linhas de distribuição de energia subterrânea. Nesse contexto, a tendência é que nos próximos anos os grandes centros urbanos migrem para os sistemas subterrâneos e isso pode se dar por vários motivos, dentre os quais a confiabilidade do sistema, a redução de acidentes, a estética das cidades, etc. Um exemplo refere-se à cidade de Porto Alegre (capital do Rio Grande do Sul), onde foi instituída a lei municipal 13.402, de 21 de março de 2023, que estabelece que as redes de infraestrutura de cabeamento para a transmissão de energia elétrica, de telefonia, de comunicação de dados via fibra óptica, de televisão a cabo e de outros cabeamentos deverão ser exclusivamente subterrâneas em um prazo máximo de quinze anos. Provavelmente outras capitais também seguirão na mesma direção, portanto, estudar os mecanismos de funcionamento dessa modalidade de rede e especialmente os problemas que as circundam é pertinente e importante.

Relativo às redes subterrâneas, neste trabalho o ponto de partida refere-se a um dos principais problemas oriundos desse sistema: as Descargas Parciais (PDs, do inglês *partial dischages*), que ocorrem no interior da isolação dos cabos, originando a deterioração elétrica. A IEC 60270 define PDs como descargas elétricas localizadas que ligam apenas parcialmente o isolamento entre os condutores e que podem ou não ocorrer adjacentes a um condutor. Já de acordo com a IEEE 400.3, PDs são pulsos rápidos que podem conter energia em frequências muito altas (acima de 100 kHz). Tanto a IEC 60270 como a IEEE 440.3 referem-se à normatização técnica aplicada à medição de descargas elétricas localizadas em meios isolantes.

As principais patologias que acarretam a ocorrência das PDs, conforme Densley (2001), são os fatores de envelhecimento (térmico, elétrico, mecânico, químico e ambiental), as rachaduras e as vedações defeituosas nas bainhas dos cabos, a entrada de umidade no dielétrico (ou isolante do cabo) e a corrosão ou vedações defeituosas, dentre outros fatores. Cabe ressaltar também que o sistema elétrico possui uma estrutura com um sistema de proteção programado para atuar sempre que condições anormais ocorram. Porém, essas descargas não são detectadas por esse sistema de proteção, e esse é um dos principais motivos que levam a faltas que causam as interrupções no fornecimento de energia aos consumidores.

Para evitar a descontinuidade do suprimento de energia ocasionado pelas descargas – que conforme Mousavi (2005) são inicialmente incipientes, mas, no decorrer do tempo, vão se tornando catastróficas –, estudos têm abordado alternativas como ferramentas de *machine learning (ML)* ou aprendizado de máquina.

Essas técnicas consistem em uma série de algoritmos que extraem informação de um conjunto de dados e buscam um padrão para realizar predições. Deisenroth *et al.* (2021) afirmam que a aprendizagem de máquina pode ser entendida como uma forma de encontrar padrões e estrutura de dados de forma automática, otimizando os parâmetros de um modelo. Além disso, esses autores afirmam que o ML baseia-se na linguagem matemática para expressar conceitos que parecem intuitivamente óbvios, mas que são surpreendentemente difíceis de formalizar. Uma vez formalizados adequadamente, é possível obter *insights* sobre a tarefa que se deseja desenvolver. Em relação aos algoritmos de ML, quando a saída assumir somente um conjunto de rótulos predeterminados, eles são chamados de algoritmos de classificação.

Os algoritmos de classificação, conforme Braga *et al.* (2016), envolvem a tarefa de atribuir a um padrão desconhecido uma entre várias classes conhecidas. A resolução de problemas de classificação caracteriza-se por aprendizado supervisionado, no qual exemplos de padrões são apresentados às entradas; já as classes correspondentes são apresentadas às saídas da rede durante o processo de aprendizado.

Já a classificação relacionada a fontes de PDs, conforme Jineeth *et al.* (2018), requer classificadores inteligentes devido à sua complexidade. Portanto, como os modelos de classificação têm sido utilizados com sucesso para o reconhecimento de PDs e outros distúrbios em pesquisas correlatas, nesta pesquisa eles estão sendo utilizados especificamente para simulações com dados reais em redes de energia elétrica subterrânea. Portanto, o objetivo deste trabalho consiste em, a partir do modelo de um dos padrões de PDs de uma rede real, investigar a capacidade de classificação das PDs em redes de distribuição subterrânea, utilizando ML. Para tanto, foram escolhidos cinco classificadores para condução das simulações: SVM, ANN, árvore de decisão, K-NN e *ensemble*.

Machine learning

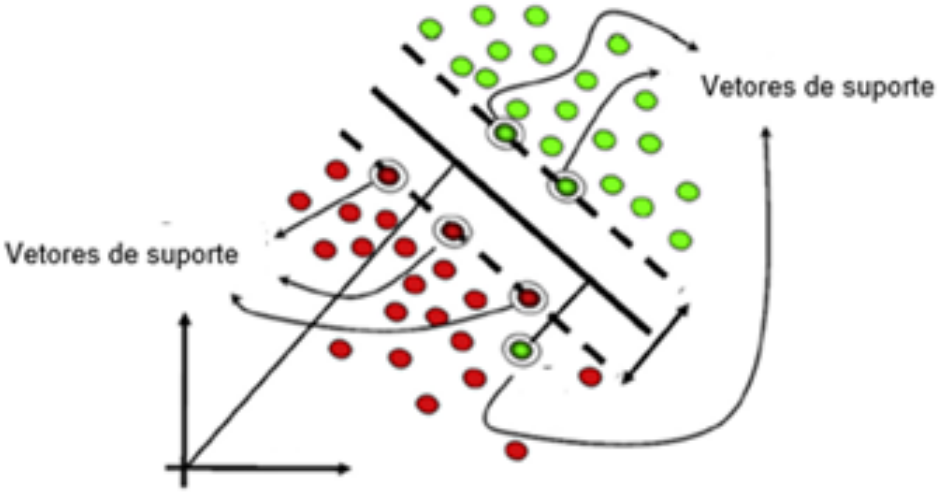
A partir de 1970, houve uma expansão no uso da inteligência artificial para solução de problemas reais. A crescente complexidade dos problemas a serem computacionalmente tratados, a velocidade e o volume de dados gerados por diferentes setores motivaram o desenvolvimento de ferramentas computacionais mais sofisticadas e autônomas. A maioria dessas ferramentas é baseada em ML, uma subárea da inteligência artificial que faz parte de várias tecnologias

Revista Vivências | Erechim | v. 21 | n. 43 | p. 61-77 | jul./agos. 2025
DOI: https://doi.org/10.31512/vivencias.v21i43.1242

atualmente utilizadas. Os algoritmos de ML podem ser divididos em algoritmos descritivos e preditivos. Os algoritmos descritivos são extraídos de padrões dos valores de determinado conjunto de dados. Já os algoritmos preditivos consistem em uma função que, dado um conjunto de exemplos rotulados, constrói um estimador. Se um domínio conhecido for um conjunto de valores nominais, tem-se um problema de classificação e o estimador gerado é um classificador. Na sequência, são apresentados os classificadores utilizados neste trabalho, e é importante observar nesse cenário, que a escolha dos classificadores usados se deu a partir da adequação dos mesmos para detecção e identificação de PDs.

O classificador SVM (do inglês *Support Vector Machines*), tal como definem Jineeth *et al.* (2018), é uma ferramenta de aprendizagem supervisionada com capacidade para lidar com problemas de classificação complexos, que envolvem aprendizagem estatística criando planos de separação para cada classe de dados. Ele pertence à categoria dos modelos de maximização de margens. Os recursos não lineares e as características de PDs são mapeados em um espaço de recursos de alta dimensão usando uma função base radial gaussiana de mapeamento não linear. No que refere às características do classificador SVM, Hastie, Tibshirani e Friedman (2008) afirmam que ele possui alto poder preditivo e também alta capacidade de extrair combinações dos dados disponíveis, porém, baixo índice de interpretabilidade. A maioria dos algoritmos SVM classifica-se apenas em duas classes (binárias), separando o conjunto de dados de treinamento e um vetor de classes. As classes são separadas por um hiperplano, conforme mostrado na Figura 1.

Figura 1: Exemplo de SVM

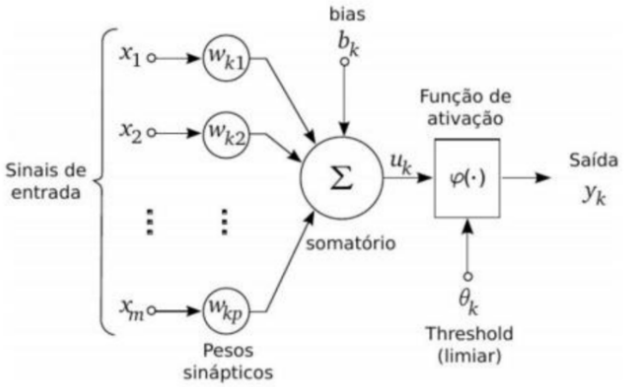


Fonte: Adaptado de Babu e Mohan (2015)

A busca por um sistema computacional cuja capacidade de processamento se aproximasse do cérebro humano motivou o surgimento das redes neurais artificiais ou ANN (do inglês *Artificial Neural Networks*) (HAYKIN, 2009). Essa é uma ferramenta comprovada para classificar um conjunto complexo de dados que apresente diferenças muito pequenas em suas características. As ANNs são sistemas computacionais distribuídos, compostos por unidades de processamento simples, densamente interconectada, conhecidas como neurônios artificiais (FACELI *et al.*, 2023). Essas unidades são interligadas por um grande número de conexões. Na maioria das arquiteturas, essas conexões, que simulam as sinapses, possuem pesos associados. Os pesos são

responsáveis por armazenar o conhecimento da rede. Já as sinapses são elos ou conexões que permitem a transmissão de informações entre neurônios. Na Figura 2 é apresentado o diagrama de neurônio que forma base para o projeto de ANN.

Figura 2: Modelo de um neurônio (HAYKIN, 2009).



As variáveis que fazem parte do modelo apresentado na Figura 2 são: os sinais de entrada, representados por $x_1, x_2, \ldots, x_n$; os pesos sinápticos do neurônio k, representados por $w_{k1}, w_{k2}, \ldots, w_{kn}$; a saída do combinador linear devido aos sinais de entrada, denotada por u_k; um termo linear que tem efeito de aumentar ou reduzir a entrada da função de ativação chamado bias e representado por b_k; a função de ativação, representada por $\varphi(.)$; e o sinal de saída do neurônio, representado por y_k. No que refere à detecção e ao diagnóstico de faltas em sistemas de energia, conforme Mas'ud *et al.* (2016), a RNA é apropriada para detecção e diagnóstico de faltas, tanto *online* como *offline*, reduzindo, com isso, a dependência de especialistas para interpretação de faltas e reduzindo também o custo e o trabalho de implementação visual.

O classificador baseado em árvore de decisão (do inglês *decision tree*), conforme Murthy (1998), é uma forma de representar regras subjacentes aos dados com estruturas hierárquicas e sequências que particionam os dados recursivamente. Conforme Faceli *et al.* (2023), esse classificador pertence à categoria de métodos simbólicos, e a ideia básica foi aperfeiçoada a partir dos métodos sugeridos por Quinlan (1986). No trabalho de Hastie, Tibshirani e Friedman (2008), eles apresentam como desvantagens desse classificador o fato de que ele possui baixo poder preditivo e apresenta baixa capacidade de extrair combinações dos dados disponíveis e médio grau de interpretabilidade. Já como vantagem citam a velocidade e a robustez contra distribuição de classes sobrepostas, especialmente contra rotulagem incorreta dos dados de treinamento. Esse classificador pode ser utilizado para descobrir se os dados contêm classes de objetos bem separadas.

O classificador K-NN (do inglês *K-Nearest-Neighbour*) ou, em português, "vizinhos mais próximos", pertence à categoria de métodos baseados em distância. É um algoritmo de ML para reconhecimento de padrões e cujo objetivo é classificado segundo a classe de maior frequência, considerando os k vizinhos mais próximos no espaço de atributos. Nesse método, é possível atribuir pesos às colaborações dos vizinhos, de modo que aqueles com maior proximidade contribuam mais para a média dos representantes. Conforme Webb e Kopsey (2011), o procedimento dos k-vizinhos mais próximos para classificar uma medida x em uma classe C consiste em determinar os valores (ou vetores) de dados de treinamento mais próximos da medição , usando a métrica de distância adequada; e, por último, atribuir x à classe com mais

Revista Vivências | Erechim | v. 21 | n. 43 | p. 61-77 | jul./agos. 2025
DOI: https://doi.org/10.31512/vivencias.v21i43.1242

representantes (ou votos) dentro do conjunto com mais valores (ou vetores) próximos. Os únicos aspectos que requerem pré-especificação são o número de vizinhos, a métrica de distância e o conjunto de dados de treinamento. Ao apresentarem as características do classificador K-NN, Hastie, Tibshirani e Friedman (2008) destacam que ele possui alto poder preditivo, capacidade média de extrair combinações dos dados disponíveis e baixa capacidade de interpretabilidade, além disso, apresenta pouca informação acerca da essência das decisões tomadas.

O classificador *ensemble*, diferentemente dos outros, utiliza-se de resultados de um conjunto de modelos preditivos, aplicados sobre a mesma base de dados para atingir melhores resultados. Ele pertence à categoria de métodos baseados em modelos múltiplos descritivos. Conforme Kuncheva e Whitaker (2003), um classificador *ensemble* costuma ser mais preciso do que qualquer um dos classificadores individuais. A abordagem adotada, conforme Webb e Copsey (2011), consiste em identificar um conjunto de candidatos a modelos, treinar os classificadores usando o conjunto de padrões rotulados e adotar o classificador que forneça o melhor desempenho de generalização. Isso resulta em um único classificador, que pode ser aplicado em todo o espaço de recursos. Em relação às suas principais características, Webb e Copsey (2011) avaliam que os esquemas de combinação dos classificadores podem ser de diferentes tipos (ANN, K-NN, árvore de decisão, etc). As combinações podem ocorrer nos diferentes níveis de saída dos componentes. Já conforme Ren *et al.* (2016), a principal teoria por trás dos métodos *ensemble* é a decomposição *bias*-variância-covariância. O *bias* mede a diferença média entre a previsão e o resultado desejado. Essa decomposição oferece justificativa teórica para melhorar o desempenho de um *ensemble* em relação aos preditores de sua base.

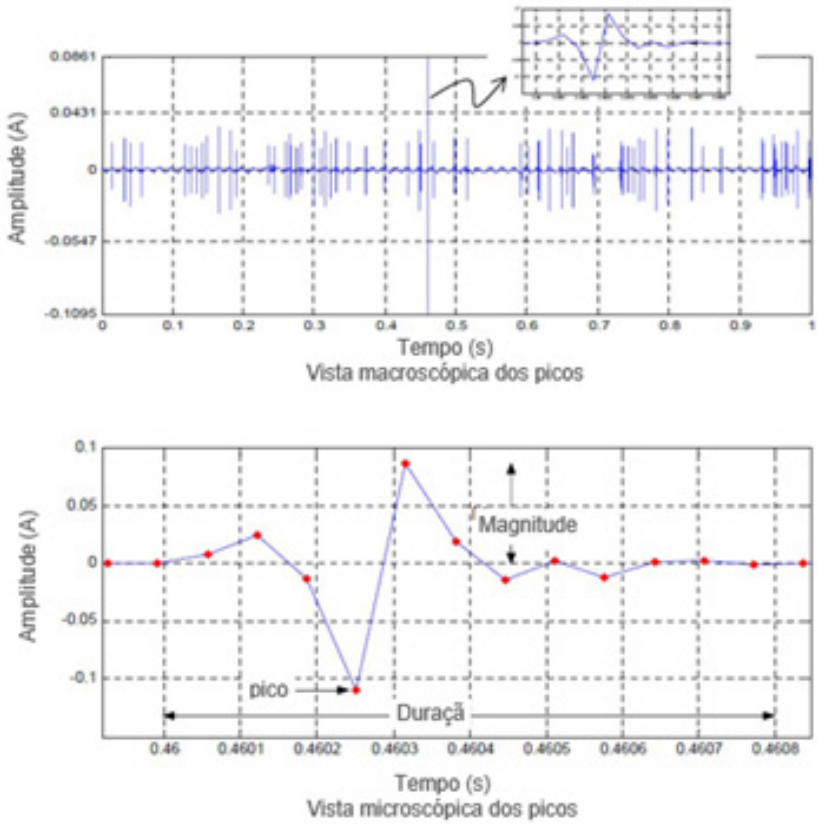
Há que se destacar, que não existe técnica universal de ML que possa resolver todo tipo de problema. Conforme Faceli *et al*. (2023), a escolha do classificador depende das características do problema que está sendo solucionado, razão pela qual sugerem que sejam escolhidos diversos candidatos a classificadores e salientam que é indicado realizar ajustes nos parâmetros dos modelos para otimizar os resultados.

Metodologia

Conforme já indicado neste trabalho, são investigados e comparados vários métodos de ML como estratégia para classificar e detectar as PDs em sistemas de distribuição de energia elétrica subterrânea. Para atingir o objetivo proposto, foram adotados procedimentos metodológicos que serão apresentados na sequência.

Para se obter o conjunto de dados com distúrbios, foi escolhido, na literatura técnica, um dos padrões de PDs embasados em dados reais. A seguir, é apresentado esse padrão característico estudado em Mousavi (2005) e que pode ser observado na Figura 3. Na parte superior do entalhe, é apresentado um modelo em que vários picos com duração e magnitude diferentes se manifestam na performance do sinal de alta frequência. Já na parte inferior da Figura 3 é mostrado o modelo de pico da descarga parcial com realce. Os picos são aumentos súbitos e temporários na tensão por ocasião do fornecimento de energia elétrica, resultando em uma tensão momentânea na linha de alimentação.

Figura 3: Um dos padrões de PDs



Fonte: Adaptado de Mousavi (2005)

Para obter esse sinal, além de experimentos controlados, Mousavi (2005) realizou uma análise de um conjunto de dados adquiridos em um circuito alimentador de média tensão subterrâneo, composto por cabo isolado em polietileno reticulado (XLPE) de uma área residencial da região de Dallas, no Texas.

Para as simulações desse trabalho, foi utilizado o software Simulink/ Matlab® e gerado um sinal de tensão com frequência de 60 Hz a partir de uma rede real do sistema radial seletivo da companhia CEEE - Grupo Equatorial, situada em Porto Alegre - RS. A esse sinal foi incorporado de forma sequencial o padrão de descarga parcial também resultante de dados reais, apresentado na Figura 3, e o resultado pode ser conferido na Figura 4. A Figura 6 apresenta dados referentes à implementação desse mesmo processo, no entanto os padrões de faltas foram incorporados de forma não sequencial. Um sinal com 5.001 amostras foi gerado para cada caso, sendo considerado como vetor de entrada (correntes) a ser utilizado nas simulações.

Foram avaliados os resultados da utilização de classificadores para identificação de PDs em redes de energia elétrica subterrâneas, tendo sido utilizados cinco classificadores: SVM, ANN, árvore de decisão, K-NN e *ensemble*.

Os dados foram simulados a partir do aplicativo "*classification learner*", integrante do software Matlab® (versão R 2021a). Nesse processo, contou-se ainda com um recurso chamado "*misclassification costs*", que considera incluir penalidades para previsões incorretas de classes. O treinamento requer um conjunto de dados de entradas e saídas conhecidas para esses dados (ou seja, classes rotuladas). Usou-se, além disso, configuração de otimização de hiperparâmetros (do

Revista Vivências | Erechim | v. 21 | n. 43 | p. 61-77 | jul./agos. 2025
DOI: https://doi.org/10.31512/vivencias.v21i43.1242

inglês *Hyperparameter Optimization*), o que pode afetar o desempenho do modelo, reduzindo o erro de classificação e melhorando sua acurácia.

Utilizou-se aprendizado supervisionado, enfatizando as comparações entre os diferentes classificadores. Conforme Pomares *et al.* (2018), durante a fase do treinamento, o algoritmo pode ser avaliado em cada etapa com a ajuda do estado atual do modelo.

Em relação à validação, prosseguem Pomares *et al.* (2018), o aplicativo oferece os seguintes esquemas integrados de validação, que indicam a acurácia preditiva calculada a partir do modelo treinado:

No validation: todos os dados de entrada são usados para treinar o modelo. A matriz de confusão é calculada usando os mesmos dados de treinamento.

Holdout validation: os dados de entrada são divididos em dois conjuntos complementares: um para treinamento e o outro para validação do modelo.

Cross-validation: essa opção seleciona conjuntos disjuntos para particionar os dados. Enquanto apenas um conjunto é utilizado para validação do modelo, os outro são utilizados para treinamento. Esse processo é repetido vezes e a matriz de confusão resultante é obtida com as médias aritméticas dos resultados de cada iteração.

As simulações para esse trabalho foram realizadas utilizando para validação o esquema *cross- validation* com cinco *fold* ou conjunto de dados *(k = 5),* uma vez que esse esquema possui uma proteção contra *overfitting,* particionando o conjunto de dados e estimando a precisão em cada *fold*. O *overfitting* ocorre quando o modelo é sobreajustado aos dados do treinamento.

Para generalizar o desempenho do modelo, o conjunto de dados de teste foi avaliado por um conjunto de métricas ou medidas de desempenho (a partir da matriz de confusão). Essa matriz foi usada para testar padrões de diferentes classificadores. Hamel (2009) destaca que, com base na matriz de confusão, é possível definir várias medidas de desempenho adicionais que são comumente usadas nas avaliações de modelos de classificação. Antes de apresentar as métricas, sugere-se observar, no Quadro 1, o modelo genérico da matriz de confusão para duas classes, onde Tp (verdadeiros positivos) representa as instâncias positivas preditas de forma correta pelo classificador, Tn (verdadeiros negativos) representa as instâncias negativas preditas corretamente pelo classificador, Fp (falsos positivos) representa as instâncias negativas preditas incorretamente pelo classificador, e Fn (falsos negativos) representa as instâncias positivas preditas incorretamente pelo classificador. Conforme Faceli *et al.* (2023), as medidas citadas podem ser facilmente generalizadas para problemas com mais de duas classes.

Quadro 1: Layout da matriz de confusão

		Verdadeiro	**Falso**
Observado	**Verdadeiro**	Verdadeiro positivo (Tp)	Falsos negativos (Fn)
	Falso	Falsos positivos (Fp)	Verdadeiro negativo (Tn)
		Predito	

Origem: Adaptado de Hamel (2009)

Em relação às taxas de erro que podem ser calculadas a partir da matriz de confusão, tem-se a taxa de erro na classe positiva (ou taxa de falsos positivos) e a taxa de erro na classe negativa (ou taxa de falsos negativos), representadas, respectivamente, pelas Equações 1 e 2:

$$TFP = \frac{Fp}{Fp + Tn} (1)$$

$$TFN = \frac{Fn}{Fn + Tp} (2)$$

Em relação à taxa de erro geral do modelo, é possível determiná-la por meio da Equação 3. Sobre isso, Hamel (2009) destaca que essa taxa é obtida a partir dos dados da matriz de confusão, já mencionada anteriormente.

$$Erro = \frac{Fp + Fn}{Fp + Fn + Tp + Tn} (3)$$

Observe-se que quanto menores os resultados dessas taxas, melhor é a performance do classificador.

Ainda com base nas informações apresentadas no Quadro 1, a teoria resultante também foi empregada para calcular as métricas de classificação que, conforme Hamel (2009), são acurácia, sensibilidade, especificidade e precisão. Nesta pesquisa, são utilizadas somente as três primeiras métricas, já que só interessam os resultados oriundos da classe minoritária. A acurácia (do inglês *accuracy*), representada pela Equação 4, mede o percentual de acertos do classificador (tanto positivos quanto negativos). A sensibilidade (do inglês *recall*), representada pela Equação 5, mede o percentual de instâncias positivas da amostra que foram classificadas corretamente pelo classificador (como positivas). Já a especificidade (do inglês *specificity)*, representada pela Equação 6, mede o percentual de instâncias negativas da amostra que foram classificadas corretamente pelo classificador (como negativas).

$$Acurácia = \frac{Tp + Tn}{Tp + Tn + Fp + Fn} \quad (4)$$

$$Sensibilidade = \frac{Tp}{Tp + Fn} (5)$$

$$Especificidade = \frac{Tn}{Tn + Fp} (6)$$

Em relação à qualidade das respostas dos classificadores, é relevante ressaltar que embora tenham sido registrados avanços tecnológicos relacionados com as ferramentas de simulação, ainda se verificam limitações de softwares associados à aplicação dessas ferramentas.

Um fator importante que se deve observar ao se trabalhar com conjunto de dados (ou volume de dados) é que na maioria das vezes não é possível aplicar algoritmos diretamente sobre os dados. Técnicas de pré-processamento são frequentemente utilizadas para corrigir problemas e tornar o conjunto de dados adequado.

Formalmente, um conjunto de dados pode ser representado por uma matriz , em que é o número de objetos e é o número de atributos de entrada de cada objeto. O valor define a dimensionalidade dos objetos ou do espaço dos objetos. Quando os valores dos atributos identificam categorias às quais os objetos pertencem, denominam-se classes, e são assumidos valores discretos . Tem-se nesse caso uma tarefa de classificação.

Revista Vivências | Erechim | v. 21 | n. 43 | p. 61-77 | jul./agos. 2025
DOI: https://doi.org/10.31512/vivencias.v21i43.1242

De acordo com Faceli *et al.* (2023), para ajudar a entender os dados e identificar as tarefas pré-processamento ou mesmo para garantir a qualidade e a facilidade da interpretação dos resultados obtidos, é importante elaborar uma caracterização desses dados e também uma exploração inicial, com estatística descritiva e/ou técnicas de visualização.

Especificamente na parte de pré-processamento, tarefas podem ser utilizadas antes de algoritmos. Dentre as principais tarefas, estão:

a. **Amostragem de dados**: relacionada ao tratamento do tamanho do conjunto de dados originados, gerando eficiência computacional. Existem basicamente três abordagens: amostragem aleatória simples, amostragem estratificada e amostragem progressiva;
b. **Dados desbalanceados ou desbalanceamento das classes de dados**: em um conjunto de dados reais, o número de objetos varia para as diferentes classes. Outro enfoque é apresentado por Cieslak e Chawla (2008), que salientam que um conjunto de dados nos quais uma classe é particularmente rara, porém mais importante (ao que se denomina de conjunto de dados desbalanceados), pode ser um problema se não for aplicado tratamento diferenciado;
c. **Limpeza de dados:** essa técnica deve ser aplicada em caso de dados ruidosos, dados inconsistentes, dados redundantes ou dados incompletos;
d. **Transformação de dados**: técnica utilizada quando os dados apresentam valores simbólicos e /ou numéricos;
e. **Redução de dimensionalidade**: técnica utilizada quando o conjunto de dados apresenta número elevado de atributos.

E sobre a importância de identificar dados ruidosos, Kumar *et al.* (2024) enfatizam que o ruído tem se tornado um desafio, especialmente ao realizar investigações sofisticadas baseadas em dados de PDs, nesse caso, o ruído pode distorcer a forma de onda, e os pulsos de ruído podem ser reconhecidos como pulsos de PDs. Isso pode facilmente resultar na detecção incorreta da atividade de PDs.

Alguns recursos são disponibilizados para resolver os problemas de deficiência nos dados. Para resolver adversidades com dados desbalanceados, é possível utilizar técnicas que procuram balancear artificialmente um conjunto de dados. As principais alternativas consistem em redefinir o tamanho do conjunto de dados e utilizar diferentes custos de classificação para diferentes classes. O aplicativo "*Classification Learner"* possui a função "*misclassification costs",* que permite especificar os custos (ou pesos) de classificação que forem julgados incorretos, antes mesmo de iniciar o treinamento dos modelos. Já para resolver problemas com dados ruidosos, algumas técnicas têm se mostrado eficientes, como análises extensivas e métodos de processamento de sinais necessários para a separação, análise do espectro de frequências, técnicas estatísticas, reconhecimento de formas de onda e padrões de análise tempo-frequência.

Os problemas mencionados são frequentes em conjuntos de dados reais, e o tratamento deles, em geral, leva a um melhor desempenho dos algoritmos de ML.Neste trabalho, quando adequado e necessário, foram utilizados os recursos indicados. Ao final das simulações, os resultados foram comparados com os achados publicados em artigos científicos recentes com enfoque para o mesmo tema.

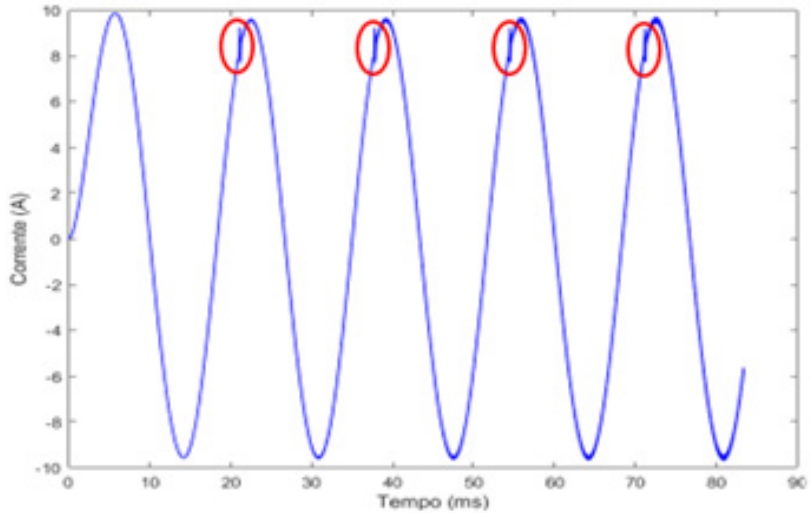
Resultados e discussões

Nesta seção, são apresentados os resultados provenientes de algoritmos executados, juntamente com as discussões das análises referentes às performances das tarefas de classificação de PDs em redes de distribuição de energia elétrica subterrânea.

Ressalta-se que a taxa de falsos positivos foi incluída nas análises, principalmente porque essa variável representa as taxas em que o classificador não conseguiu identificar a ocorrência de distúrbios (no caso de PDs). É importante salientar também que o classificador SVM, apesar de não ter apresentado boa performance, foi incluído nas análises por ser especial para classificação binária, por apresentar rapidez de processamento no caso de dados próximos e por ser menos sensível aos ruídos.

Na Figura 4, inicialmente, é apresentado o sinal de corrente com destaque para a ocorrência das PDs. É possível identificar os distúrbios ocorrendo em vários pontos de forma sequencial, em um período de aproximadamente 85 ms.

Figura 4: Sinal de corrente com a presença de quatro PDs.



Fonte: o autor.

A intensão aqui foi utilizar um sistema que pudesse classificar as PDs assim que elas ocorrem. Esse sistema é necessário porque a quantidade de dados gerados a cada segundo quando o sistema de energia está operando é expressiva, dificultando, com isso, a identificação das PDs.

Portanto, para a obtenção de parâmetros para comparação de resultados, foram utilizados cinco classificadores. A escolha desses classificadores se deu em função da presença dessas modalidades em alguns trabalhos correlatos. Também porque os resultados desses cinco classificadores foram os que melhor responderam às métricas utilizadas nas análises. Na Tabela 1 são apresentados os resultados das taxas de erro e métricas. As taxas de falsos positivos (TFP) representam a proporção de instâncias negativas preditas incorretamente pelo classificador (no caso deste trabalho, as taxas em que foi classificada incorretamente a presença das PDs). Já as (taxas de falsos negativos) TFN representam as instâncias positivas preditas incorretamente pelo classificador (no caso deste trabalho, as taxas em que o classificador identificou erroneamente a operação normal do sistema).

Revista Vivências | Erechim | v. 21 | n. 43 | p. 61-77 | jul./agos. 2025
DOI: https://doi.org/10.31512/vivencias.v21i43.1242

Aqui cabe observar que, em relação à precisão do classificador, quanto menores as taxas de erro e/ou quanto mais próximas da unidade ficarem as métricas, melhor será a precisão do classificador.

Tabela 1: Taxas de erro e performance dos classificadores (Quatro PDs)

Classificador	Taxas de erro			Métricas		
	Falsos negativos	Falsos positivos	Erro geral	Acurá-cia	Sensibi-lidade	Especifi-cidade
SVM	0,1586	0,6190	0,1664	0,8636	0,8427	0,3809
ANN	0,0034	0,3213	0,0088	0,9912	0,9965	0,6786
Árvore decisão	0,0004	0,0952	0,0020	0,9980	0,9995	0,9048
K-NN	0,0022	0,0476	0,0030	0,9976	0,9978	0,9524
Ensemble	0,0006	0,0119	0,0008	0,9993	0,9994	0,9881

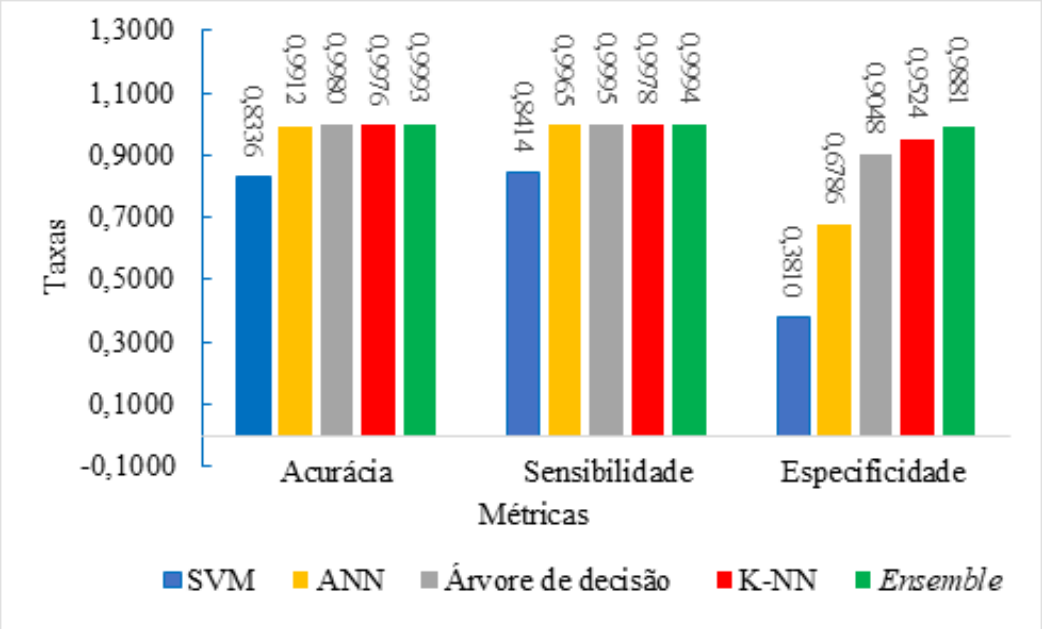
Fonte: O autor.

Ao observar a Tabela 1, verifica-se que o classificador *ensemble* apresentou as melhores taxas, portanto, teve o melhor desempenho, tanto em relação à taxa de falsos negativos (0,0006ou 0,06%) quanto no que concerne à taxa de falsos positivos (0,0119 ou 1,19%e à taxa de erro geral (0,0008 ou 0,08%). Já o classificador SVM apresentou taxas com o desempenho inferior. Nos classificadores árvore de decisão e K-NN, houve um aumento da taxa de falsos negativos: no primeiro, de 0,0010 ou 0,1%, e no segundo de 0,0022 ou 0,22%, uma vez que este último apresentou maior índice de erro ao identificar as amostras em que o sistema operou normalmente. Já nos falsos positivos, ocorreu o inverso, pois o classificador árvore de decisão errou mais ao tentar identificar as PDs.

Em relação à qualidade dos classificadores para predição, cabe salientar que embora seja uma métrica importante, a acurácia define apenas a assertividade do classificador de forma geral. Já as taxas de especificidade, essas sim definem a proporção de assertividade do classificador quanto à identificação das PDs (classe minoritária). E as taxas de sensibilidade aqui definem com exatidão a operação normal do sistema.

Ainda referente à Tabela 1, os dados nela apresentados evidenciam que o classificador *ensemble* apresentou a melhor performance. A taxa de 0,9881 (ou 98,81%) corresponde à especificidade indica que foi o classificador *ensemble* que identificou as PDs com a maior taxa de assertividade. Já o classificador SVM apresentou a menor taxa de assertividade em todas as métricas e pode ser conferido na Figura 5 a seguir, na qual foram representadas graficamente as métricas de todos os classificadores analisados.

Figura 5: Gráfico representativo da performance dos classificadores (quatro PDs).

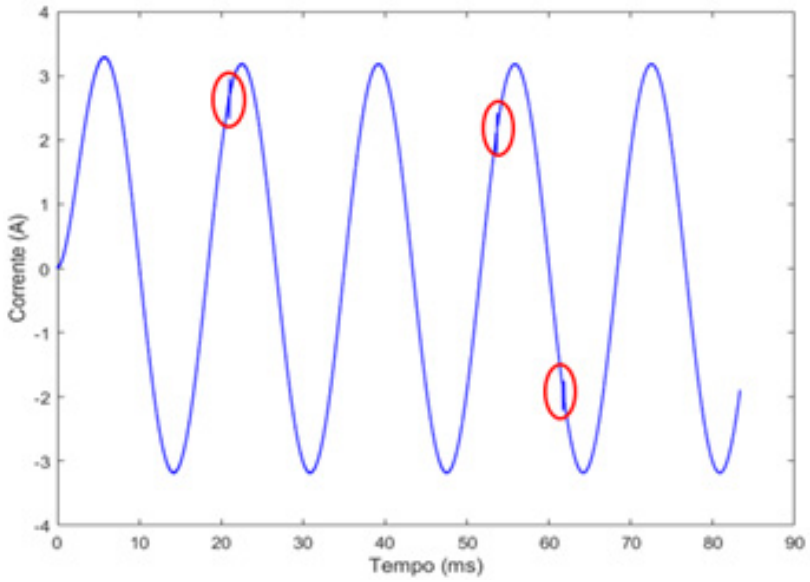


Fonte: O autor.

Outra avaliação realizada foi referente à taxa do erro geral e acurácia (Tabela 1). Como a acurácia é o resultado da diferença entre a unidade e a taxa de erro geral (acurácia = 1 – erro geral), é natural que no classificador árvore de decisão a acurácia (0,9960 ou 99,60%) seja menor do que no classificador K-NN (0,9970 ou 99,70%). E o inverso ocorre em relação à taxa do erro geral, de modo que ela é menor no classificador K-NN (0,0030 ou 0,30%) e maior no classificador árvore de decisão (0,0040 ou 0,4%).

Em outro cenário, foram realizadas novas simulações para identificar as PDs em outro cenário: quando as PDs não aparecem de forma não sequencial, tal como apresentado na Figura 6.

Figura 6: Sinal de corrente com a presença de três PDs.



Fonte: O autor.

Revista Vivências | Erechim | v. 21 | n. 43 | p. 61-77 | jul./agos. 2025
DOI: https://doi.org/10.31512/vivencias.v21i43.1242

A Tabela 2 reúne os resultados para taxa de erro e métricas, porém, agora, referente aos dados que geraram a Figura 6.

Tabela 2: Taxas de erro e performance dos classificadores (Três PDs)

Classificador	Taxas de erro			Métricas		
	Falsos negativos	Falsos positivos	Erro geral	Acurácia	Sensibilidade	Especificidade
SVM	0,0059	0,0159	0,0060	0,9940	0,9941	0,9841
ANN	0,0022	0,1746	0,0044	0,9956	0,9978	0,8254
Árvore decisão	0,0635	0,0008	0,0016	0,9984	0,9992	0,9365
K-NN	0,0030	0	0,0030	0,9971	0,9970	1.0000
Ensemble	0,0022	0	0,0022	0,9978	0,9978	1.0000

Fonte: O autor.

Na Tabela 2, ao avaliar-se as taxas de erro, observa-se que os resultados foram melhores quando comparados com o conjunto de dados utilizados para gerar a Tabela 1. Ao analisarem-se as métricas, constata-se que elas tiveram resultados melhores especialmente quando observadas as taxas de falsos positivos (TFP) que identificam a presença das PDs sendo classificadas incorretamente como operação normal do sistema. Para esse caso, houve dois classificadores (K-NN e *ensemble*) que não tiveram erros de classificação para a presença de PDs (ou falsos positivos). Também observou-se que os outros três classificadores apresentaram taxas de falsos positivos menores ao comparar com resultados da Tabela 1. Assim, ao avaliar as principais métricas, evidencia-se que, para o caso da especificidade, se comparado com os resultados da Tabela 1, todos os classificadores apresentaram melhores resultados. Inclusive os classificadores K-NN e *ensemble* atingiram valor 1 ou 100% na métrica especificidade, o que denota que ambos foram precisos ao identificar as PDs.

Outros trabalhos que abordam o tema foram consultados e comparados. No trabalho de Sahoo, Karmakar e Panigrahy (2020), foram usados recursos de ML para avaliar a condição de saúde do isolamento de cabos a partir da magnitude da descarga. Os resultados das simulações dos classificadores utilizados foram semelhantes ao encontrado neste trabalho, com exceção do classificador SVM, que apresentou resultado superior. A justificativa para essa diferença pode ser em função de que o trabalho foi desenvolvido a partir de cabos de alta tensão e não de média tensão. Os classificadores árvore de decisão e *ensemble* não foram avaliados pelos autores. No trabalho de Saleh *et al.* (2022), os autores recorreram ao ML para detecção e classificação de PDs com o intuito de avaliar a confiabilidade do sistema de isolamento de cabos. Um dos classificadores que os autores empregaram no trabalho foi árvore de decisão, e o resultado também foi similar ao encontrado na presente pesquisa.

Considerações finais

A identificação das PDs em sistemas de energia subterrâneo é importante para que ocorra a manutenção – preventiva ou corretiva –, evitando, com isso, acidentes graves ou falta de energia. Portanto, a proposta deste trabalho é, a partir do modelo de um dos padrões de PD de uma rede real, investigar a capacidade de classificação de padrões de PDs em redes de distribuição de energia elétrica subterrânea, utilizando o ML.

O resultado da classificação quando da utilização do aplicativo *classification learner* (Matlab®) forneceu resultados importantes, mas principalmente a partir dele foi possível comparar os resultados de performance apresentados pelos diversos classificadores de PDs utilizados nas simulações. Foram elaboradas simulações com cinco classificadores com o objetivo de avaliar a capacidade de separação e de classificação dos sinais de PDs. Os resultados foram considerados eficazes, ao serem comparados com trabalhos correlatos.

O classificador *ensemble,* em primeiro lugar, e o K-NN, em segundo, foram os classificadores que apresentaram as menores taxas de erro e métricas nas simulações de dois bancos de dados diferentes (com quatro e três PDs). As análises levaram em conta primeiramente as taxas de falsos positivos (TFP) que identificam a presença de PDs, sendo classificadas incorretamente como operação normal do sistema; e também as taxas de especificidade, que identificaram com precisão a presença de PDs. As outras métricas também foram consideradas para completar a análise, no entanto, em uma escala secundária.

Apesar de este trabalho se encontrar em fase inicial, já foi possível comparar resultados obtidos com resultados da literatura técnica. Os resultados obtidos indicam claramente que a abordagem proposta possui validade e relevância como técnica de classificação de PDs, contudo, os autores destacam a necessidade de testar outras modalidades de PDs para se chegar a uma interpretação mais detalhada e robusta sobre o comportamento delas.

Referências

BABU, N. R.; MOHAN, B. J. Fault classification in power systems using EMD and SVM. **Ain Shams Engineering Journal**, v. 8, set.2015. Disponível em: https://core.ac.uk/download/pdf/81105226.pdf. Acesso em: 08 out.2023.

BRAGA *et al*. **Redes neurais artificiais**: teoria e aplicações. Rio de Janeiro: LTC, 2016.

CIESLAK, D. A.; CHAWLA, N.V. Start Globally, Optimize Locally, Predict Globally: Improving Performance on Imbalanced Data. **Eighth IEEE International Conference on Data Mining**, p. 143-152. Doi: 10.1109/ICDM.2008. Disponível em: https://ieeexplore.ieee.org/document/4781109. Acesso em 08 jun. 2025.

DEISENROTH, M. P. *et al*. **Mathematics for Machine Learning**. Cambridge: Cambridge University Press, 2011.

Revista Vivências | Erechim | v. 21 | n. 43 | p. 61-77 | jul./agos. 2025
DOI: https://doi.org/10.31512/vivencias.v21i43.1242

DENSLEY, J. Ageing mechanisms and diagnostics for power cables - an overview, **IEEE Electrical Insulation Magazine,** v. 17 (1), p. 14-22, 2001. Disponível em: https://ieeexplore.ieee.org/document/901613. Acesso em: 12 set. 2023.

FACELI, K. *et al.* **Inteligência artificial** - uma abordagem de aprendizado de máquina. 2. ed. Rio de Janeiro: LTC, 2023.

HAMEL, L. **Knowledge Discovery with Support Vector Machines**, p. 231-235, [*S. l.: s. n.*], out.2009. Disponível em: https://onlinelibrary.wiley.com/doi/book/10.1002/9780470503065. Acesso em: 23 set. 2023.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning**: Data mining, inference, and prediction. Springer Verlang: Nova Iorque, 2008.

HAYKIN, S. **Neural networks and learning machines**. 3rd ed. New York: Prentice Hall, 2009.

INTERNACIONAL ELECTROTECHNICAL COMISSION. **IEC 60270**: High voltage test technique - partial discharge measurements. IEC: Switzerland, 2000.

INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS. **IEEE 400.3:** Guide for PD Testing of Shielded Power Cable Systems in a Field Environment. IEEE: New York, USA, 2006.

JINEETH, J. *et al*. Classification of Partial Discharge Sources in XLPE Cables by Artificial Neural Networks and Support Vector Machine. **IEEE Electrical Insulation Conference (EIC**), p. 407-411, 2018. Disponível em: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9548051. Acesso em: 23 set. 2023.

KUMAR, H.; SHAFIQ, M.; KAUHANIEMI, K.; ELMUSRATI, M. A. Review on the Classification of Partial Discharges in Medium-Voltage Cables: Detection, Feature Extraction, Artificial Intelligence-Based Classification, and Optimization Techniques. **Energies**, v.17, p. 1-31, fevereiro, 2024. Doi: 10.3390/en17051142. Disponível em: https://doi.org/10.3390/en17051142. Acesso em: 09 Jun. 2025.

KUNCHEVA L. I.; WHITAKER, C. J. Measures of Diversity in Classifier Ensembles and Their relationship with the Ensemble Accuracy. **Machine Learning**, v. 51, n. 2, p. 181-207, 2003. Disponível em: https://link.springer.com/content/pdf/10.1023/a:1022859003006.pdf. Acesso em 23 set. 2023.

MAS'UD, A. A. *et al*. Artificial neural network application for partial discharge recognition: Survey and future directions**. Energies**, v. 9, n. 8, 2016. Disponível em: https://www.mdpi.com/1996-1073/9/8/574. Acesso em 10 jun. 2025.

MOUSAVI, M. J. **Underground distribution cable incipient fault diagnosis system**. 2005. 195 f. Ph.D. thesis (Electrical Engineering). Texas AM University, 2005. Disponível em: https://oaktrust.library.tamu.edu/handle/1969.1/4675. Acesso em 09 ago. 2023.

MURTHY, S. K. Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. **Data Mining and Knowledge Discovery**, v. 2, p. 345-389, 1998. Disponível em: https://cs.nyu.edu/~roweis/csc2515-2006/readings/murthy_dt.pdf. Acesso em 23 set. 2023.

POMARES, A. *et al.* Ground Extraction from 3D Lidar Point Clouds with the Classification Learner App, ***26th Mediterranean Conference on Control and Automation (MED)***, Zadar, Croatia, p. 1-9, 2018. Disponível em: https://riuma.uma.es/xmlui/bitstream/handle/10630/16062/MED%202018.pdf?sequence=1&isAllowed=y. Acesso em: 21 ago. 2023.

QUINLAN, J.R. Induction of decision trees. **Machine Learning Journal**, v.1, p. 81–106, 1986. Disponível em: https://hunch.net/~coms-4771/quinlan.pdf. Acesso em: 20 out. 2023.

REN, Y. *et al*. Ensemble Classification and Regression-Recent Developments, Applications and Future Directions [Review Article], ***IEEE Computational Intelligence Magazine***, v. 11, n. 1, p. 41-53, 02. 2016. Disponível em: https://ieeexplore.ieee.org/document/7379058. Acesso em 08 out. 2023.

SAHOO, R.; KARMAKAR, S.; PANIGRAHY, S. Health Index Analysis of XLPE Cable Insulation using Machine Learning Technique. **IEEE 7th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)**, p. 1-6, 2020. Doi: 10.1109/UPCON50219.2020.9376573. Disponível em: https://ieeexplore.ieee.org/document/9376573. Acesso em: 08 jun. 2025.

SALEH, M. A. *et al*. Detection and Classification of Defects in XLPE Power Cable Insulation via Machine Learning Algorithms. **3rd International Conference on Smart Grid and Renewable Energy (SGRE)**, p. 1-6, 2022. Disponível em: https://ieeexplore.ieee.org/document/9774113. Acesso em 10 jun. 2025

WEBB, A. R.; Kopsey, K. D. **Statistical pattern recognition**. 3 ed. Hoboken: Wiley, 2011.