Auditable Inference in Physics-Regulated Classifiers: *PhysicsOnly*/*NeuralOnly*/Hybrid Modes and a Fuzzy Reading of the Differentiable Physics Layer

Douglas de Tarso da Silva[1], **Eduardo Crestana Guardia**[1], **Roberto Akira Yamachita**[1], **Gabriel Antônio Fanelli de Souza**[1]

[1]Programa de Pós-Graduação em Engenharia Elétrica (PPGEE)
Universidade Federal de Itajubá (UNIFEI) – Itajubá, MG – Brasil

eng.douglas.tarso.dts@outlook.com, eduardo.guardia@unifei.edu.br

akira@unifei.edu.br, fanelli@unifei.edu.br

Abstract. *Threshold-guided AI classifiers increasingly embed an engineering rule (e.g. harmonic-distortion thresholds) as a differentiable physics layer, but their verdicts are usually delivered as a single opaque label. We contribute an* auditable inference protocol *for such models: from one forward pass we expose three verdicts per norm* —PhysicsOnly, NeuralOnly*, and* Hybrid— *with the trained mixing weight α, a near-boundary flag, and a stream-disagreement signal. We further show that the soft thresholds of the differentiable physics layer are, formally,* fuzzy membership functions*: the internal "Compliant" class becomes a graded membership, the severity scores form a near unit-sum fuzzy partition built from product t-norms (measured deviation $\leq 6.1 \times 10^{-6}$), and the α-mix acts as a measured neuro-fuzzy gate. Preserved output from a historical four-threshold power-quality artifact (195 row-level cases, seed 42) illustrates the protocol: the neural stream disagrees with the rule in 9.2–22.1% of samples, yet the hybrid contradicts the rule in only 2 samples — both within 0.008 percentage points of the PRODIST threshold; in the preserved logit mixture, $\alpha = 0.015$ accompanies the inversion of a near-tied physics verdict. A second preserved, exploratory output on 923 real AIRI substation oscillograms exhibits the same observational pattern: hybrid–rule flips stay confined to the boundary (3 windows), while latent neural–rule disagreement reaches 87–88%. These are descriptive counts for that historical artifact: current provenance auditing does not establish a source/event/view-disjoint split or normative/metrological validity, and the reproduction routes therefore fail closed. No accuracy, generalisation, drift-detection, or compliance claim follows from this case study.*

Resumo. *Classificadores de IA guiados por limiares vêm embutindo regras de engenharia (p.ex. limiares de distorção harmônica) numa camada física diferenciável, mas seus vereditos costumam sair como um único rótulo opaco. Contribuímos com um* protocolo de inferência auditável*: de um único* forward*, expomos três vereditos por norma* —PhysicsOnly, NeuralOnly *e* Híbrido— *junto do peso de mistura treinado α, de uma marca de proximidade de fronteira e de um sinal de discordância entre fluxos. Mostramos ainda que os limiares suaves da camada física diferenciável são, formalmente,* funções de pertinência fuzzy*: a classe interna "Conforme" vira pertinência graduada, os escores*

de severidade formam uma partição fuzzy de soma quase unitária construída com t-normas produto (desvio medido $\leq 6{,}1 \times 10^{-6}$*) e o* α*-mix age como* gate *neuro-fuzzy mensurado. Saídas preservadas de um artefato histórico de qualidade de energia com quatro limiares (195 casos separados por linha, semente 42) ilustram o protocolo: o fluxo neural discorda da regra em* $9{,}2$–$22{,}1\%$ *das amostras, mas o híbrido contraria a regra em apenas 2 amostras — ambas a* $0{,}008$ *ponto percentual do limite PRODIST, onde o* $\alpha = 0{,}015$ *treinado acompanha, na mistura de* logits *preservada, a inversão de um veredito físico quase empatado. Uma segunda saída histórica, exploratória, em 923 oscilografias reais AIRI de subestação exibe o mesmo padrão observacional: os flips híbrido–regra ficam confinados à fronteira (3 janelas), enquanto a discordância latente rede–regra chega a* 87–88%*. Essas contagens descrevem somente o artefato histórico: a auditoria atual de proveniência não demonstra split independente por fonte/evento/view nem validade normativa/metrológica, e as rotas de reprodução falham fechadas. O estudo não autoriza claims de acurácia, generalização, detecção de deriva ou conformidade.*

1. Introdução

Classificadores guiados por física podem embutir, de forma diferenciável, uma comparação simplificada com limiares de engenharia — p.ex. de distorção harmônica total (THD) em QEE, domínio deste artigo. Uma arquitetura representativa combina um *backbone* compartilhado (CNN 1D, BiLSTM, Transformer) com cabeças paralelas por norma, cada uma acoplada a uma *camada física diferenciável* (DPL) que computa uma classe segundo o limiar configurado e a um peso treinável α que mistura ramo aprendido e ramo físico [Silva et al. 2026b].

Para o leitor de computação: QEE caracteriza a forma de onda de tensão; a THD relaciona as componentes harmônicas à fundamental da rede (tipicamente 50 ou 60 Hz). IEEE 519, IEC 61000-2-2, o PRODIST/ANEEL (Módulo 8) e referências CIGRE têm tipologias, faixas de tensão e procedimentos de agregação distintos. Logo, comparar uma janela isolada a um número não produz, por si, conformidade normativa. O objeto do artigo é justificar a decomposição do veredito interno.

O ganho de embutir a física, contudo, vem com um custo de *governança*: na prática o modelo entrega um único rótulo (“Conforme”, “Precário”, “Crítico”, “Grave”), sem dizer *quanto* do veredito veio da regra e quanto veio da rede, nem *quão perto* o sinal está da fronteira de decisão — decomposição sem a qual o rótulo é difícil de auditar em domínio regulado. A literatura de XAI insiste que, em alto risco, modelos interpretáveis por construção são preferíveis a explicações *post-hoc* de caixas-pretas [Rudin 2019]; este artigo segue essa linha, explorando estrutura que o modelo já possui e descarta na inferência.

Este artigo contribui com dois instrumentos complementares de explicabilidade, extraídos do modelo *sem re-treino*: **(i) três modos de inferência** fatiados de um único *forward* — *PhysicsOnly* (só a regra, $\alpha = 0$), *NeuralOnly* (só a rede, $\alpha = 1$) e *Híbrido* (mistura treinada) —, cada um com veredito por norma exposto com α, proximidade de fronteira e discordância entre fluxos (Seções 3 e 5); **(ii) uma leitura *fuzzy*** da DPL: o limiar suave é função de pertinência, os escores de classe são composições por t-norma

produto (a conjunção graduada da lógica *fuzzy*; aqui, o produto) formando partição de soma quase unitária (desvio máximo medido $6{,}1 \times 10^{-6}$) e o α-mix é um *gate* neuro-fuzzy — o peso treinável que funde regra e rede — cujo valor treinado é em si diagnóstico (Seção 4). Os dois instrumentos são exercitados em um estudo de caso histórico descritivo e em uma aplicação exploratória preservada em 923 oscilografias AIRI de subestação (Seções 6 e 6.4).

Delimitamos desde já o escopo: os rótulos de treino derivam da mesma regra de THD embutida na DPL, tornando a acurácia circular [Silva et al. 2026a]; além disso, a partição histórica não tem independência por grupo comprovada. A contribuição é formal e arquitetural — interpretabilidade do artefato —, não validação preditiva, normativa ou operacional (Seção 8).

2. Trabalhos Relacionados

Redes físico-informadas. Restringir o aprendizado por leis físicas remonta às *physics-informed neural networks* (PINNs) [Raissi et al. 2019], hoje parte de uma agenda ampla de aprendizado de máquina guiado por física [Karniadakis et al. 2021]. Na taxonomia de aprendizado informado de [von Rueden et al. 2021], as PINNs integram o conhecimento prévio pela função de perda; a DPL aqui estudada o integra pela *arquitetura*, pois a regra do protótipo não é uma EDP, e sim uma comparação de limiar (THD $<$ limite) computada no próprio *forward* — instanciada em [Silva et al. 2026b] com *backbone* compartilhado, cabeças por norma e α-mix treinável.

Explicabilidade *post-hoc* vs. por construção. Métodos de atribuição como LIME [Ribeiro et al. 2016] e SHAP [Lundberg and Lee 2017] explicam *quais entradas* pesaram num veredito, e surveys organizam essas técnicas e seus riscos [Guidotti et al. 2018, Barredo Arrieta et al. 2020]. Há o argumento de que, em alto risco, explicações *post-hoc* podem ser infiéis [Rudin 2019] — e até manipuláveis adversarialmente [Slack et al. 2020]. Nosso protocolo segue a via estrutural com outro alvo: não explica *quais entradas* pesaram, e sim expõe as saídas dos dois ramos, o coeficiente α e a distância da fronteira. Os campos são fiéis aos tensores internos do grafo; sua interpretação causal ou utilidade para usuários não é garantida [Miller 2019].

Sistemas *fuzzy* e neuro-*fuzzy*. A leitura graduada da classe interna "Conforme" conecta a DPL ao arcabouço de conjuntos *fuzzy* [Zadeh 1965]: o limiar suave é uma função de pertinência e a composição de classes usa t-normas (Seção 4). Sistemas de inferência *fuzzy* clássicos constroem o veredito por regras linguísticas (Mamdani [Mamdani and Assilian 1975]) ou por consequentes funcionais (Takagi–Sugeno [Takagi and Sugeno 1985]); o neuro-*fuzzy* ANFIS [Jang 1993] treina por gradiente as pertinências. A DPL é o caso dual: pertinências fixadas pela norma, treinando-se só o *gate* α por norma, que medimos.

Explicabilidade em QEE (2023–2026). Trabalhos recentes estudam SHAP sobre LSTM [Jrhilifa et al. 2025], explicações por espaço latente [Machlev et al. 2023], classificação auditável e transferência de domínio com EBM [Pacheco et al. 2026], restrições físicas na detecção de afundamentos de tensão [Shi et al. 2025] e detecção OOD em QEE [Zhang et al. 2026]. Esse levantamento narrativo, verificado por DOI e metadados editoriais, não é uma RSL nem sustenta exaustividade. Nosso recorte é a decomposição explícita regra–rede do próprio artefato e de seu peso de fusão.

3. O Modelo-Objeto: DPL Multi-Norma e os Três Modos

Esta seção descreve o modelo *como implementado*, pois o protocolo depende dos detalhes. O DiffNormPQ [Silva et al. 2026b] contém quatro limiares históricos rotulados IEEE 519 (τ=8,0%), CIGRE TB 261 (*planning level*, 6,0%), IEC 61000-2-2 (nível de compatibilidade, 8,0%) e PRODIST Módulo 8 (10,0%). Esses números têm escopos distintos; aqui são apenas a configuração do artefato e não um parecer de conformidade. Cada cabeça n emite uma classe interna pela razão $r = x/\tau_n$ do THD x ao limite: *Conforme* ($r < 1$), *Precário* ($1 \le r < 1{,}5$), *Crítico* ($1{,}5 \le r < 2$) e *Grave* ($r \ge 2$).

3.1. THD diferenciável

A DPL computa o THD diretamente do vetor de atributos harmônicos normalizados $\mathbf{h} = (h_1, \ldots, h_{50})$, com h_1 a componente fundamental:

$$\mathrm{THD}(\mathbf{h}) \;=\; \frac{\sqrt{\sum_{j=2}^{50} h_j^2}}{\max(h_1,\, \epsilon)} \times 100, \tag{1}$$

com $\epsilon = 10^{-6}$ para estabilidade. Essa é a fórmula interna H2–H50 do artefato em forma diferenciável; ela não substitui os contratos de medição de cada referência. O mesmo cálculo gera o rótulo e integra o grafo de computação — fato central para a circularidade.

3.2. Limiar suave configurado

Para um limite τ sobre o THD x, a DPL substitui a indicadora não-diferenciável $\mathbf{1}[x < \tau]$ pelo limiar suave

$$\mu_\tau(x) \;=\; \sigma\big(k\,(\tau - x)\big) \;=\; \frac{1}{1 + e^{-k(\tau - x)}}, \tag{2}$$

onde σ é a sigmoide e $k > 0$ a inclinação (operador `soft_below` da implementação). No *checkpoint* histórico, $k = 4{,}0$: μ atinge $\sigma(4) \approx 0{,}982$ a 1 ponto percentual (pp) do limite. Esse hiperparâmetro não foi calibrado contra incerteza metrológica. A *zona de transição* de uma norma é o intervalo em que $\mu_\tau \in (0{,}01,\, 0{,}99)$, de meia-largura $\ln(99)/k \approx 1{,}15$ pp para $k = 4$.

3.3. Bandas de severidade por limiar

Cada cabeça física instancia a DPL com as três fronteiras $(b_1, b_2, b_3) = (\tau_n,\, 1{,}5\,\tau_n,\, 2\,\tau_n)$ e compõe os escores de classe por produtos de pertinências e complementos:

$$\begin{aligned} s_0(x) &= \mu_{b_1}(x), & s_1(x) &= \mu_{b_2}(x)\,\big(1 - \mu_{b_1}(x)\big), \\ s_2(x) &= \mu_{b_3}(x)\,\big(1 - \mu_{b_2}(x)\big), & s_3(x) &= 1 - \mu_{b_3}(x), \end{aligned} \tag{3}$$

com $s_0, \ldots, s_3$ associados a Conforme, Precário, Crítico e Grave, e *logits* físicos $\ell^{\text{phy}} = 4\,\mathbf{s}$ (o fator 4 apenas dá amplitude ao *softmax*). Note: s_1 codifica "abaixo de $1{,}5\tau$ **e não** abaixo de τ" — conjunção graduada formalizada na Seção 4.

3.4. Mistura treinável e os três modos

Sejam ℓ_n^{neu} os *logits* da cabeça neural da norma n (saída de uma MLP sobre o *embedding* do *backbone*) e ℓ_n^{phy} os *logits* físicos da Eq. (3). A saída por norma é a combinação convexa

$$\ell_n \;=\; \alpha_n\,\ell_n^{\text{neu}} \;+\; (1-\alpha_n)\,\ell_n^{\text{phy}}, \qquad \alpha_n = \sigma(\rho_n) \in (0,1), \tag{4}$$

com ρ_n um parâmetro livre, treinado a partir de $\alpha_n{=}0{,}5$ (inicialização neutra). A Eq. (4) torna os três modos um *fatiamento* do mesmo *forward*: *PhysicsOnly* avalia ℓ_n^{phy} (equivale a $\alpha_n = 0$) — o *baseline determinístico* dos limiares configurados; *NeuralOnly* avalia ℓ_n^{neu} ($\alpha_n = 1$) — detecta se a rede aprendeu algo além da regra; *Híbrido* avalia ℓ_n com o α_n treinado — a saída padrão do artefato histórico. O custo de auditoria é marginal: uma passagem, três vereditos por norma.

4. Leitura Fuzzy da Camada Física

4.1. O limiar suave é uma função de pertinência

A função μ_τ da Eq. (2) é uma função de pertinência do conjunto *fuzzy* "x está conforme o limite τ" [Zadeh 1965]: contínua, estritamente decrescente, com $\mu_\tau(\tau) = \frac{1}{2}$ e limites 1 ($x \ll \tau$) e 0 ($x \gg \tau$); $k \to \infty$ recupera o conjunto clássico (indicadora), e k finito codifica uma transição suave em torno do limiar; neste artefato, essa transição não foi calibrada contra incerteza metrológica. A Figura 1 mostra as pertinências dos quatro limiares configurados (IEEE 519 e IEC 61000-2-2 compartilham $\tau = 8\%$).

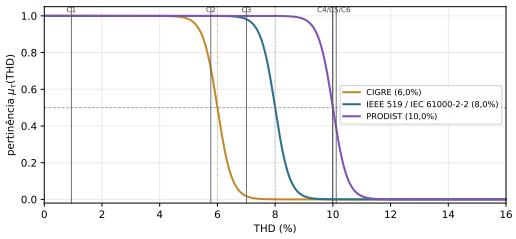


Figura 1. Pertinências μ_τ do artefato ($k = 4{,}0$), limiares configurados; verticais: THD de C1–C6 (Tab. 3); tracejado: $\mu = 0{,}5$.

4.2. As bandas de severidade são uma partição fuzzy

Os escores da Eq. (3) são composições *fuzzy*: usando a t-norma produto $T_P(a,b) = ab$ e a negação padrão $N(a) = 1 - a$,

$$s_1 = T_P\big(\mu_{b_2},\, N(\mu_{b_1})\big), \qquad s_2 = T_P\big(\mu_{b_3},\, N(\mu_{b_2})\big), \qquad s_3 = N(\mu_{b_3}), \tag{5}$$

ou seja, "Precário" é o grau de *(abaixo de* $1{,}5\tau$*)* $\wedge$ *(não abaixo de* τ*)*, e analogamente para as demais. Da Eq. (3) segue a identidade

$$\sum_{c=0}^{3} s_c(x) - 1 \;=\; \mu_{b_1}(x)\big(1-\mu_{b_2}(x)\big) \;+\; \mu_{b_2}(x)\big(1-\mu_{b_3}(x)\big) \;\geq\; 0, \tag{6}$$

que quantifica o quanto a família $\{s_c\}$ se afasta de uma partição de Ruspini (soma unitária) [Ruspini 1969]: o desvio só é não-nulo quando duas fronteiras consecutivas "vazam" uma na outra, e decai com $e^{-k\,\Delta b}$ no espaçamento $\Delta b = \tau_n/2$ entre fronteiras. Verificada numericamente para os quatro limiares (grade de THD de 0 a 30%, passo 10^{-3} pp), a Eq. (6) dá desvio máximo $6{,}1 \times 10^{-6}$ (CIGRE, fronteiras mais próximas: 6, 9 e 12%) e $2{,}1 \times 10^{-9}$ (PRODIST: 10, 15 e 20%). Portanto, na grade examinada, a família é quase unitária; ela não é probabilidade calibrada nem valida, por si, um sistema de inferência fuzzy.

4.3. Agregação multi-norma como cardinalidade fuzzy

No nível agregado, a implementação combina os quatro limiares com a *contagem suave*

$$C(x) \;=\; \sum_{n=1}^{4} \mu_{\tau_n}(x) \;\in\; [0, 4], \tag{7}$$

que pode ser lida como cardinalidade *fuzzy* (σ-contagem) dos limiares satisfeitos; a variante diferenciável mapeia C a classes por gaussianas centradas em 4, 3, $1{,}5$ e 0 (Conforme atende às 4 normas; Precário, 3; Crítico, 1–2; Grave, 0), e o protocolo reporta também a contagem dura $C_{\mathrm{hard}} = \#\{n : x < \tau_n\}$. Na Tabela 3, longe das fronteiras $C \approx C_{\mathrm{hard}}$; na zona de transição divergem informativamente (p.ex. $C = 3{,}71$ para um sinal que atende às 4 normas, uma delas por só $0{,}23$ pp); a Tabela 5 registra a ocupação dessa zona no output AIRI.

4.4. O α-mix como *gate* neuro-fuzzy mensurado

Na Eq. (4), o veredito final pondera *logits* físicos (ℓ^{phy}) e evidência aprendida (ℓ^{neu}); na terminologia neuro-*fuzzy* [Jang 1993], α_n é o peso de agregação, treinado por gradiente, entre um bloco de inferência *fuzzy* fixado na configuração e um bloco conexionista. Diferente do ANFIS — que treina as pertinências —, aqui elas permanecem fixas e o grau de liberdade da fusão é α_n. Como a mistura ocorre antes do *softmax* e os ramos podem ter escalas distintas, α_n é coeficiente de mistura dos *logits*, não fração causal do veredito. No *checkpoint* histórico, $\alpha = 0{,}0131/0{,}0107/0{,}0131/0{,}0150$ (IEEE 519/CIGRE/IEC/PRODIST) — $1-\alpha$ multiplica os *logits* físicos, mas não mede sua contribuição causal. A leitura *fuzzy* dá a semântica descritiva: em vez de "Crítico", reportar "pertence $0{,}71$ a conforme-CIGRE e $0{,}9999$ a conforme-IEEE 519" (valores reais do caso C2 da Tabela 3).

5. Protocolo de Auditoria

Sobre os três modos e a leitura *fuzzy*, o protocolo define o *contrato de auditoria*: o conjunto mínimo de metadados que toda inferência deve emitir, por norma, para ser justificável *a posteriori* (Tabela 1, campos como implementados).

No nível agregado, o protocolo emite ainda o THD computado, a contagem C_{hard} e o veredito de pior caso entre normas. Três campos merecem justificativa:

α_n **por cabeça** expõe o coeficiente escalar da mistura de *logits*. Valores próximos de zero são compatíveis com deferência no snapshot, mas não decompõem causalmente a decisão nem provam que rótulos circulares sempre levam $\alpha_n \to 0$ [Silva et al. 2026a].

Tabela 1. Contrato de auditoria emitido por cabeça a cada inferência; nenhum campo exige re-treino — todos derivam do *forward* único e das Eqs. (1)–(4).

Campo	Semântica
`severity_class`	classe decidida pelo modo corrente (com índice e *softmax* completo)
`confidence_pct`	escore máximo de *softmax*, não calibrado como confiança
`alpha`	peso α_n efetivo do modo (0 físico; 1 neural; treinado no híbrido)
`physics_class`	veredito *PhysicsOnly*, sempre preenchido, em qualquer modo
`neural_class`	veredito *NeuralOnly*, sempre preenchido, em qualquer modo
`stream_disagreement`	verdadeiro sse `physics_class` $\neq$ `neural_class`
`near_boundary`	verdadeiro sse $\lvert\text{THD} - \tau_n\rvert < 0{,}5$ pp
`distance_to_boundary_pp`	margem assinada $\text{THD} - \tau_n$ (positiva = acima do limiar)
`norm_limit_thd_pct`	o limite τ_n aplicado, lido da própria DPL

near_boundary é verdadeiro quando o THD está a menos de 0,5 pp do limite — região de $\mu_{\tau_n} \approx 0{,}5$. A tolerância (configurável, mas não calibrada metrologicamente) é deliberadamente mais estrita que a meia-largura de $1{,}15$ pp da zona *fuzzy* de $k = 4$. ***stream_disagreement*** é verdadeiro quando *PhysicsOnly* difere de *NeuralOnly*; expõe divergência entre ramos, sem identificar o que a rede aprendeu ou qual ramo está correto.

Como política a validar, casos com *near_boundary* ou híbrido≠física podem ser encaminhados à revisão humana. O estudo não autoriza que os demais sigam automaticamente.

6. Estudo de Caso

6.1. Configuração e estado da reprodutibilidade

Auditamos as saídas históricas preservadas do classificador de [Silva et al. 2026b] para o *checkpoint* `DiffNormPQ_v2_real_data_masked_meta.pt` (SHA-256 `05bfaa0c...`) e uma partição por linha de 195 casos, extraída de um corpus convertido de 971 linhas, com semente 42 e tolerância de fronteira 0,5 pp. Os JSONs preservados são aritmeticamente consistentes entre si, mas a ordem histórica índice→WAV não foi selada e a independência por fonte/evento/view não foi demonstrada. Por isso, os drivers atuais bloqueiam antes de carregar dados ou *checkpoint*; as tabelas abaixo descrevem o snapshot histórico, não uma reprodução corrente.

O snapshot também registra acurácia macro interna de $99{,}74\%$ (três cabeças $100{,}0\%$; PRODIST $98{,}97\%$). Esse número não é evidência preditiva: combina rótulos circulares, split não qualificado e forte desbalanceamento (134–150 de 195 casos na classe Grave).

6.2. Auditoria agregada por norma

A Tabela 2 resume os pesos α_n, a discordância entre fluxos e a proximidade de fronteira.

O output preservado ilustra a contradição interna: o ramo neural discorda da regra em 9,23–22,05% das amostras, mas o *híbrido* treinado quase sempre coincide com a física (0,00% em três normas; 1,03% no PRODIST), ao mesmo tempo em que

Tabela 2. Auditoria histórica por limiar ($n = 195$): divergência interna de fluxos (Física≠Neural), saída operacional vs. regra (Híbrido≠Física), fração com $|\mathrm{THD} - \tau_n| < 0{,}5$ pp (Fronteira) e fração na zona *fuzzy* $\mu \in (0{,}01;\ 0{,}99)$ (Transição).

Rótulo	α_n	Física≠Neural	Híbrido≠Física	Fronteira	Transição
IEEE 519	0,0131	12,31%	0,00%	0,00%	3,59%
CIGRE	0,0107	22,05%	0,00%	0,51%	3,59%
IEC 61000-2-2	0,0131	11,79%	0,00%	0,00%	3,59%
PRODIST	0,0150	9,23%	1,03%	1,54%	2,05%

$\alpha_n \in [0{,}0107;\ 0{,}0150]$. Como as escalas dos *logits* podem diferir, essa coexistência não identifica causa. A métrica útil não é "ganho de acurácia", e sim expor a discordância latente, a fronteira e o peso real da rede.

6.3. Casos históricos auditados

A Tabela 3 materializa o protocolo em oito casos do snapshot: todas as linhas marcadas como fronteira (C2, C4–C6), todos os flips híbrido≠física (C4, C5) e quatro exemplares (C1, C3, C7, C8).

Tabela 3. Casos auditados no snapshot histórico (seleção determinística, semente 42). $\mu_6/\mu_8/\mu_{10}$: pertinência a conforme-CIGRE/IEEE 519–IEC/PRODIST; C/C_h: contagem de conformidade suave (Eq. 7) e dura; F/N/H: pior caso (máxima severidade) agregado entre as quatro normas em *PhysicsOnly*/*NeuralOnly*/Híbrido (C=Conforme, P=Precário, K=Crítico, G=Grave). Os flips de C4–C5 ocorrem na cabeça PRODIST (Tabela 4); dominados por CIGRE (K) no agregado, não aparecem nesta coluna.

Caso	Origem	THD (%)	μ_6	μ_8	μ_{10}	C / C_h	F/N/H	Seleção
C1	SEED_PQ	0,94	1,000	1,000	1,000	4,00 / 4	C/P/C	exemplar
C2	PLAID	5,77	0,714	1,000	1,000	3,71 / 4	C/K/C	fronteira
C3	PLAID	7,01	0,017	0,981	1,000	2,98 / 3	P/K/P	exemplar
C4	PQ_Class.1	9,992	0,000	0,000	0,508	0,51 / 1	K/K/K	front.+flip
C5	PQ_Class.1	9,994	0,000	0,000	0,506	0,51 / 1	K/K/K	front.+flip
C6	SEED_PQ	10,12	0,000	0,000	0,385	0,39 / 0	K/G/K	fronteira
C7	PQ_Class.1	24,97	0,000	0,000	0,000	0,00 / 0	G/G/G	exemplar
C8	PQD_16C	20 549	0,000	0,000	0,000	0,00 / 0	G/G/G	exemplar

Três leituras concentram o valor do protocolo:

(i) Discordância latente mascarada pelo *gate*. Em C1 (THD $0{,}94\%$, folga de 5 a 9 pp), o *NeuralOnly* diz Precário em três normas enquanto regra e híbrido dizem Conforme; em C2, o neural diz Crítico para CIGRE. Sem o protocolo essa divergência seria invisível: no snapshot, o híbrido fica próximo da física; com os ramos expostos, a discordância vira item de auditoria, sem afirmar o que a rede de fato aprendeu.

(ii) Vereditos graduados entre normas. C3 (THD $7{,}01\%$) é o caso textual da leitura *fuzzy*: pertence $0{,}017$ a conforme-CIGRE e $0{,}981$ a conforme-IEEE 519 — Precário numa norma, Conforme nas demais ($C = 2{,}98$). O rótulo único esconderia que a não-conformidade é específica da norma mais estrita.

(iii) Os dois desvios registrados no snapshot estão na fronteira. C4 e C5 (THD $9{,}9921\%$ e $9{,}9940\%$) estão a $0{,}008$ e $0{,}006$ pp do limite PRODIST: $\mu_{10} \approx 0{,}507$, empate quase perfeito. A Tabela 4 decompõe: a física, quase empatada ($0{,}454$ vs. $0{,}427$), diz Conforme; a neural, Precário com $0{,}997$; com $\alpha = 0{,}0150$, o híbrido inverte para Precário ($0{,}476$ vs. $0{,}464$). São as duas divergências do híbrido em relação ao rótulo duro no snapshot PRODIST, pois o rótulo, derivado da regra ($\mathrm{THD} < 10$), é Conforme. Com α pequeno, o nos dois registros preservados, a mistura acompanha *logits* físicos quase empatados e *near_boundary* verdadeiro. O protocolo localiza e decompõe aritmeticamente esses casos, sem atribuir causalidade isolada a α.

Tabela 4. Flip híbrido≠física no caso C4 (PRODIST, THD $9{,}9921\%$, margem $-0{,}008$ pp): probabilidades *softmax* por modo (C5 é quase idêntico, $\mu_{10} = 0{,}506$). O rótulo da regra dura é Conforme; o híbrido erra por $0{,}012$.

Modo	p(Conforme)	p(Precário)	p(Crítico)	p(Grave)
PhysicsOnly (α=0)	$\mathbf{0{,}454}$	0,427	0,060	0,060
NeuralOnly (α=1)	0,003	$\mathbf{0{,}997}$	0,000	0,000
Híbrido (α=0,0150)	0,464	$\mathbf{0{,}476}$	0,033	0,028

Por fim, a coluna Transição da Tabela 2 mede a ocupação da zona *fuzzy* de cada norma ($\mu \in (0{,}01;\ 0{,}99)$): o restante do corpus está saturado ($\mu \approx 0$ ou 1), coerente com a distribuição de THD da validação (mediana 951%, máximo $20\,549\%$; ver Seção 8). Em leitura conservadora: neste snapshot a fração de decisões sensíveis aos limiares é pequena, e nela se concentram os desvios híbrido–regra registrados.

6.4. Aplicação exploratória ao AIRI

O artefato histórico `results_repro_real.json` registra a aplicação dos três modos, sem ajuste, a 923 oscilografias AIRI [Evdakov et al. 2026] (tensão de barramento, fase A). Com $f_s = 1600$ Hz e f_0 estimada, o maior harmônico observável variou de H12 a H17 (732/923 em H16); portanto não há suporte H2–H50 completo. Os rótulos AIRI representam criticidade de evento, não severidade harmônica, e não fornecem oráculo de acurácia ou conformidade. A Tabela 5 é apenas decomposição exploratória preservada.

Tabela 5. Output exploratório histórico: 923 oscilografias AIRI, mesmos modos e *checkpoint*. Conf. física: fração rotulada Conforme no *PhysicsOnly*; Híbrido≠Física: contagem; demais colunas como na Tabela 2.

Norma	Conf. física	Física≠Neural	Híbrido≠Física	Fronteira	Transição
IEEE 519	86,67%	87,65%	2	2,06%	2,93%
CIGRE	83,97%	88,08%	1	1,41%	3,47%
IEC 61000-2-2	86,67%	87,65%	2	2,06%	2,93%
PRODIST	88,73%	14,95%	0	0,87%	1,52%

Três padrões descritivos concentram a leitura. **(i)** O híbrido contraria a física em 3 janelas ($\leq 0{,}22\%$ por norma), *todas* com *near_boundary* verdadeiro ($\mu \in [0{,}52;\ 0{,}55]$); a coexistência observacional de α pequeno e empate físico reaparece, sem isolar causalidade. **(ii)** A discordância física–neural é 87–88% em três cabeças (e $15{,}0\%$ no PRODIST). Isso motiva um candidato a indicador de deriva, mas sem calibração, baseline independente ou oráculo OOD não valida um detector. **(iii)** Criticidade e severidade dissociam-se

sob o modelo de limiares: 61 das 100 faltas são rotuladas Conforme nas quatro cabeças no modo físico (60 no híbrido; THD mediano de faltas $4{,}59\%$), e as faltas ocupam mais alguma zona de transição *fuzzy* (35% contra $4{,}2\%$ dos registros sem evento). Essas associações não têm inferência populacional.

7. Discussão

Escopo declarado: circularidade por construção. Com $\alpha_n \approx 0{,}01$, o híbrido coincide quase sempre com o ramo físico no snapshot, mas o valor do coeficiente, isoladamente, não quantifica a contribuição neural porque as escalas dos *logits* podem diferir. Os $99{,}74\%$ internos são circulares — não devem ser lidos como generalização [Silva et al. 2026a]. O AIRI não constitui validação externa: seus rótulos de criticidade não são um oráculo para severidade harmônica, e o checkpoint permanece P27-não-qualificado. O valor demonstrado do protocolo é tornar a mistura interna *visível e auditável* — α exposto, modos separados, fronteira marcada —, fornecendo rastros estruturados para inspeção; isso não prova causalidade nem utilidade humana [Holzinger et al. 2019].

O que o modo *NeuralOnly* mede. A discordância física–neural de 9–22% (Tabela 2), com viés de maior severidade (C1–C2), é compatível com uma cabeça neural divergente da regra, mas não identifica a causa. Monitorar *stream_disagreement* é um candidato barato a indicador de deriva; limiar, falsos alarmes e poder de detecção ainda precisam ser calibrados em dados independentes.

Custo e implantação. Toda a auditoria deriva de um único *forward* com retenção dos fluxos intermediários: sem segundo modelo nem amostragem de perturbações (como em atribuição [Ribeiro et al. 2016]); pós-processamento $O(N)$ por cabeça. Arquiteturalmente, o contrato pode ser emitido no ciclo perceber–classificar–explicar–registrar, e o modo *PhysicsOnly* pode servir de comparação. Isso ainda não foi qualificado em hardware-alvo nem autoriza *fallback* operacional ou ensaio de conformidade.

Transferibilidade. A decomposição pode ser adaptada a outros modelos que misturem regra diferenciável e ramo aprendido. A semântica do limiar, a incerteza e o comportamento de α dependem de cada domínio; não inferimos convergência, conformidade ou utilidade externa sem novo experimento.

8. Ameaças à Validade

Listamos as ameaças sem atenuantes.

Circularidade definicional. Os rótulos por norma derivam da regra de THD que a própria DPL implementa: nenhum número de acurácia aqui é evidência de aprendizado generalizável. É a ameaça dominante — e o objeto que o protocolo expõe.

Proveniência e split P27. O corpus histórico não sela grupo-fonte, evento, view nem a ordem índice→WAV. Uma reconstrução atual encontrou identidades e grupos cruzando a partição, mas não prova a alocação temporal original. O auditor termina `FAIL_CLOSED_CONTAMINATED_OR_UNQUALIFIED`, e os drivers bloqueiam antes de I/O. Assim, 195/923 e todas as taxas deste artigo são descrições de JSONs históricos, não evidência científica reproduzida.

Amostra e semente únicas. A partição histórica tem 195 linhas e uma semente

(42); 2 flips nessa partição e 3 no AIRI são contagens pequenas, sem inferência populacional.

Representatividade do corpus. Os cinco conjuntos públicos convertidos misturam tensão e corrente de equipamentos, com THD recomputado chegando a $20\,549\%$ (mediana 951%) — muito além do regime esperado de tensão de rede; 96–98% das amostras saturam a leitura *fuzzy*, logo as taxas de fronteira tendem a *subestimar* as de uma campanha real de tensão BT, e as classes são desbalanceadas (Grave: 69–77%).

Escopo normativo e licenças. IEEE 519, IEC 61000-2-2, PRODIST e CIGRE não são quatro normas intercambiáveis: incluem *planning level*, nível de compatibilidade e métricas agregadas, com faixas de tensão distintas. As comparações instantâneas são telas diagnósticas do protótipo, não vereditos normativos. A linhagem/licença por linha do corpus convertido não foi selada; os dados de terceiros não integram este pacote.

Hiperparâmetros do protocolo. A inclinação $k = 4{,}0$ e a tolerância de fronteira de 0,5 pp foram herdadas da implementação, sem varredura de sensibilidade: contagens de fronteira e de transição dependem monotonicamente delas e devem ser fixadas em qualquer reprodução. O nível de tensão foi fixado em BT, pois o vetor de atributos não o carrega por amostra.

Validade da leitura *fuzzy*. A formalização é uma reinterpretação matemática exata da implementação, mas não comparamos com um FIS clássico (Mamdani [Mamdani and Assilian 1975]/TSK [Takagi and Sugeno 1985]) na mesma tarefa, nem fizemos o estudo com usuários que a avaliação rigorosa de interpretabilidade exige [Doshi-Velez and Kim 2017, Miller 2019]: a utilidade XAI é argumentativa.

Escopo AIRI. A Seção 6.4 usa um canal de tensão por oscilografia e suporte harmônico variável H12–H17, abaixo do H2–H50 do modelo. Os rótulos de evento não validam severidade harmônica, detecção OOD ou conformidade; o output tampouco pode ser reexecutado enquanto P27 estiver aberto.

9. Conclusão

Definimos uma decomposição auditável e graduada de um classificador físico-guiado, sem re-treino, por três modos fatiados de um *forward* e por uma leitura *fuzzy* da DPL — pertinências de limiar, partição de severidade por t-norma produto e α-mix como *gate* neuro-fuzzy mensurado. Os JSONs históricos ilustram 9–22% de discordância interna e flips junto aos limiares; o output AIRI registra 87–88% em três cabeças. Esses números não confirmam generalização, deriva ou conformidade: circularidade, split P27-não-qualificado e suporte harmônico incompleto impedem tal leitura. O próximo gate é re-treino group-aware sobre corpus metrologicamente qualificado, seguido de reprodução independente e estudo com usuários.

Agradecimentos

Ao PPGEE/IESTI–UNIFEI. Ferramentas de IA generativa auxiliaram na organização dos artefatos históricos e na revisão editorial; os autores verificaram o conteúdo e assumem integral responsabilidade.

Referências

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.

Evdakov, A., Filatova, G., Yablokov, A., Kovalenko, A., Skachkov, E., and Makarov, I. (2026). A dataset of real-world oscillograms from electrical power grids. *Scientific Data*.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., and Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*.

Jang, J.-S. R. (1993). ANFIS: adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems, Man, and Cybernetics*.

Jrhilifa, I., Ouadi, H., Rafia, H., Jilbab, A., Mounir, N., and Lahrarti, Y. (2025). Explainable deep learning for voltage disturbance classification in microgrid AC bus using LSTM and SHAP. In *2025 13th International Conference on Systems and Control (ICSC)*.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., and Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*.

Lundberg, S. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. arXiv:1705.07874.

Machlev, R., Perl, M., Caciularu, A., Belikov, J., Levy, K. Y., and Levron, Y. (2023). Explaining the decisions of power quality disturbance classifiers using latent space features. *International Journal of Electrical Power & Energy Systems*, 148.

Mamdani, E. H. and Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man-Machine Studies*.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*.

Pacheco, E. J., Pedroso, C. M., and Kuiava, R. (2026). Auditable PQD classification and domain transfer analysis with explainable boosting machines. *IEEE Access*.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*.

Ruspini, E. H. (1969). A new approach to clustering. *Information and Control*.

Shi, L., Cui, Q., Zhang, Y., Wu, H., Gao, J., and Li, W. (2025). A voltage sag detection framework based on constraint-driven neural networks. *International Journal of Electrical Power & Energy Systems*, 173.

Silva, D. d. T., Guardia, E. C., Yamachita, R. A., and Souza, G. A. F. d. (2026a). The circularity trap in power-quality severity classification. Manuscrito companheiro em preparação.

Silva, D. d. T., Guardia, E. C., Yamachita, R. A., and Souza, G. A. F. d. (2026b). DiffNormPQ: A multi-norm physics-informed network for power-quality severity classification. Manuscrito companheiro em preparação.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). Fooling LIME and SHAP. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

Takagi, T. and Sugeno, M. (1985). Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics*.

von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., Kirsch, B., Walczak, M., et al. (2021). Informed machine learning – a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Transactions on Knowledge and Data Engineering*.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8(3):338–353.

Zhang, Y., Wang, H., Luo, B., Zheng, Y., Shahidehpour, M., Chen, W., and Lin, K. (2026). Detection of out-of-distribution power quality disturbance using dual-channel Gramian angular field with enhanced Mahalanobis distance. *IEEE Transactions on Smart Grid*.