

Mathematical and Physical Foundations of Dissolved Gas Analysis for Power Transformer Fault Diagnosis and Insulation Life Estimation

Moma Munga Ferdinand¹; Lumanji Mbunga Luc²; Mavungu Mayela Alain³; Tshikudi Tshikudi Dennis⁴

^{1;2;3;4}Institut Supérieur de Techniques Appliquées (ISTA), Kinshasa, Democratic Republic of the Congo

Publication Date: 2026/10/13

Abstract: Dissolved gas analysis (DGA) is the reference method for assessing the internal condition of oil-filled power transformers, and machine learning is increasingly used to interpret it. This article sets out, with consistent notation, the chain of concepts that links the physics of gas formation to the output of a supervised classifier and to an estimate of insulation life. It first formalises the thermochemical origin of fault gases (bond scission, Arrhenius kinetics, gas-oil partition) and the thermal ageing of cellulose (hot-spot temperature, ageing acceleration factor, loss of life). It then derives the four classifiers most often compared on DGA data, namely multinomial logistic regression, k-nearest neighbours, support vector machines and random forests, together with stacked generalisation and the measures used to evaluate them. Numerical illustrations show that the ageing acceleration factor rises from 1 at a hot-spot temperature of 110 °C to 2.71 at 120 °C, which shortens the reference insulation life from 20.5 to 7.6 years, and that averaging four trees leaves 25% to 70% of the single-tree variance depending on tree correlation. The framework is applied in a companion article.

Keywords: Dissolved Gas Analysis, Power Transformer, Arrhenius Law, Hot-Spot Temperature, Random Forest, Support Vector Machine, Stacking, Predictive Maintenance.

How to Cite: Moma Munga Ferdinand; Lumanji Mbunga Luc; Mavungu Mayela Alain; Tshikudi Tshikudi Dennis (2026) Mathematical and Physical Foundations of Dissolved Gas Analysis for Power Transformer Fault Diagnosis and Insulation Life Estimation. *International Journal of Innovative Science and Research Technology*, 11(10), 448-453. <https://doi.org/10.38124/ijisrt/26oct413>

I. INTRODUCTION

A power transformer concentrates a large share of the capital cost of a substation, and its failure interrupts supply to every load downstream. Its insulation system combines mineral oil and cellulose (paper and pressboard). Under electrical or thermal stress both materials decompose and release gases that dissolve in the oil. Sampling the oil and measuring these gases, dissolved gas analysis (DGA), reveals an incipient fault while the unit is still in service (International Electrotechnical Commission [IEC], 2022). Classical interpretation relies on key gases, ratios and graphical methods such as the Duval triangle; these are simple and widely used, but they can leave cases undiagnosed or give conflicting verdicts (Nanfak et al., 2024).

Supervised learning has therefore been applied to DGA for more than a decade. Reviews describe neural networks, support vector machines, decision trees and ensembles as the main families, and note that reported accuracy depends strongly on the data set, on class balance and on parameter tuning (Acikgoz et al., 2025; Li, 2023;

Zhang et al., 2022). On 4,580 samples from five utilities, nonlinear support vector machines and ensembles performed best (Rao et al., 2021). Within predictive maintenance, such models are one element of a chain that goes from sensing to decision (Achouch et al., 2022).

The purpose of this article is conceptual. It states the physical laws that explain why gas concentrations carry diagnostic information, then the mathematical definition of each classifier, so that the choice of model and the reading of its errors can be argued from first principles. Because the reference window requested for this work is 2021–2026, classical results (Arrhenius kinetics, the loading-guide ageing equations, and the original formulations of the classifiers) are cited through recent standards and review articles that restate them.

II. PHYSICAL MODEL

➤ Origin of Fault Gases

Mineral oil is a mixture of hydrocarbon molecules built from CH₃, CH₂ and CH groups linked by carbon–carbon bonds. A fault supplies the energy needed to break

C–H and C–C bonds; the radicals recombine into hydrogen (H_2), methane, ethane, ethylene (C_2H_4) and acetylene (C_2H_2). The energy required increases with the order of the bond formed: about 338 kJ/mol for C–H, 607 kJ/mol for C–C, 720 kJ/mol for C=C and 960 kJ/mol for C≡C (IEC, 2022). Hydrogen therefore appears under low-energy stress such as partial discharge, ethylene becomes dominant above about 500 °C, and acetylene requires the 800–1,200 °C reached in arcs. Cellulose is less stable than oil: chain scission becomes significant above about 105 °C and produces mainly carbon monoxide (CO) and carbon dioxide.

The temperature dependence of each reaction follows the Arrhenius law:

$$k_i(T) = A_i \exp\left(-\frac{E_{a,i}}{RT}\right) \quad (1)$$

Where k_i is the rate constant of the reaction forming gas i , A_i a pre-exponential factor, $E_{a,i}$ the activation energy, $R = 8.314 \text{ J/(mol}\cdot\text{K)}$ and T the absolute temperature. Figure 1 plots Equation 1 for three activation energies chosen to illustrate the ordering described above. With these values, the rate at 500 °C relative to 1,200 °C is 1.3×10^{-3} for hydrogen, 7.3×10^{-6} for ethylene and 4.5×10^{-9} for acetylene. The gas mixture is thus a thermometer of the fault: the ratio between two gases depends on local temperature, which is the physical basis of every ratio method and of the separability exploited by classifiers.

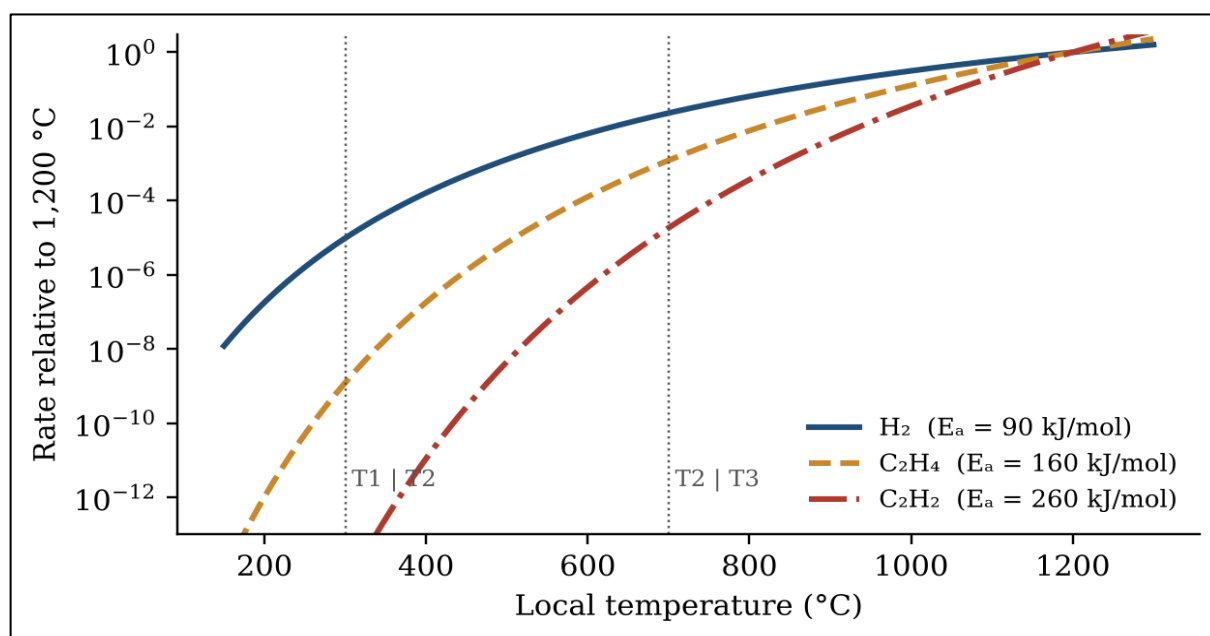


Fig 1 Relative Formation Rate of Three Fault Gases as a Function of Local Temperature

- *Note.* Equation 1 normalised at 1,200 °C. The activation energies are illustrative assumptions, not measured values. Dotted lines mark the 300 °C and 700 °C limits between thermal fault classes T1, T2 and T3.

➤ Gas–Oil Partition

What a sensor measures is the concentration dissolved in oil, C_i^{oil} , which at equilibrium is proportional to the concentration in a gas phase in contact with it:

$$C_i^{oil} = k_{O,i} C_i^{gas} \quad (2)$$

Where $k_{O,i}$ is the Ostwald coefficient. At 25 °C in mineral oil it is about 0.05 for H_2 , 0.12 for CO, 1.20 for C_2H_2 and 1.70 for C_2H_4 (IEC, 2022). For the same dissolved concentration of 100 $\mu\text{L/L}$, the equilibrium gas-phase concentration is thus 2,000 $\mu\text{L/L}$ for hydrogen but 83 $\mu\text{L/L}$ for acetylene. Hydrogen escapes most readily to a gas space or a membrane, which is one reason it is the usual target of single-gas on-line monitors, and Equation 2 must be applied before readings from gas-phase and oil-phase instruments are compared.

➤ Thermal Ageing of Cellulose

The mechanical strength of paper decreases irreversibly with time and temperature; experiments on kraft paper aged in oil at 120 °C and 160 °C show faster loss of tensile strength at the higher temperature (Duanaputri et al., 2023). The governing variable is the winding hot-spot temperature.

$$\theta_h = \theta_a + \Delta\theta_{to} + \Delta\theta_g \quad (3)$$

Where θ_a is the ambient temperature, $\Delta\theta_{to}$ the top-oil rise over ambient and $\Delta\theta_g$ the hot-spot rise over top oil. Applying Equation 1 to cellulose degradation with a reference of 110 °C (383 K) gives the ageing acceleration factor of the loading guides, and the older doubling rule gives the relative ageing rate V (Leu et al., 2023):

$$F_{AA} = \exp\left(\frac{15,000}{383} - \frac{15,000}{\theta_h + 273}\right), \quad V = 2^{(\theta_h - 98)/6} \quad (4)$$

The constant 15,000 K is E_a/R for cellulose. The two forms are consistent: differentiating Equation 4 shows that F_{AA} doubles for a rise of $\ln 2 (383)^2 / 15,000 = 6.8 \text{ K}$ near

the reference, close to the 6 K of the rule. Over an interval $[t_1, t_2]$ the equivalent ageing, the percentage loss of life and the expected life at constant temperature are:

$$L_{eq} = \int_{t_1}^{t_2} F_{AA} dt, \quad \text{LOL}(\%) = \frac{100 L_{eq}}{L_N}, \quad L(\theta_h) = \frac{L_N}{F_{AA}(\theta_h)} \quad (5)$$

With $L_N = 180,000$ h (20.5 years) the normal insulation life at 110 °C. Figure 2 shows both relations. At 120 °C, $F_{AA} = 2.71$ and $L = 66,447$ h (7.6 years); at 130

°C, $F_{AA} = 6.98$ and $L = 2.9$ years; one day at 120 °C consumes 65 h of life, or 0.036%. Two caveats apply. Equation 4 assumes that ageing is negligible well below the reference, whereas measurable ageing has been reported below a 90 °C hot spot (Leu et al., 2023), and the rate also depends on moisture and oxygen (Bilyaz et al., 2024). Chemical markers such as carbon oxides, furans and methanol give an independent view of the state of the solid insulation (Behjat et al., 2021).

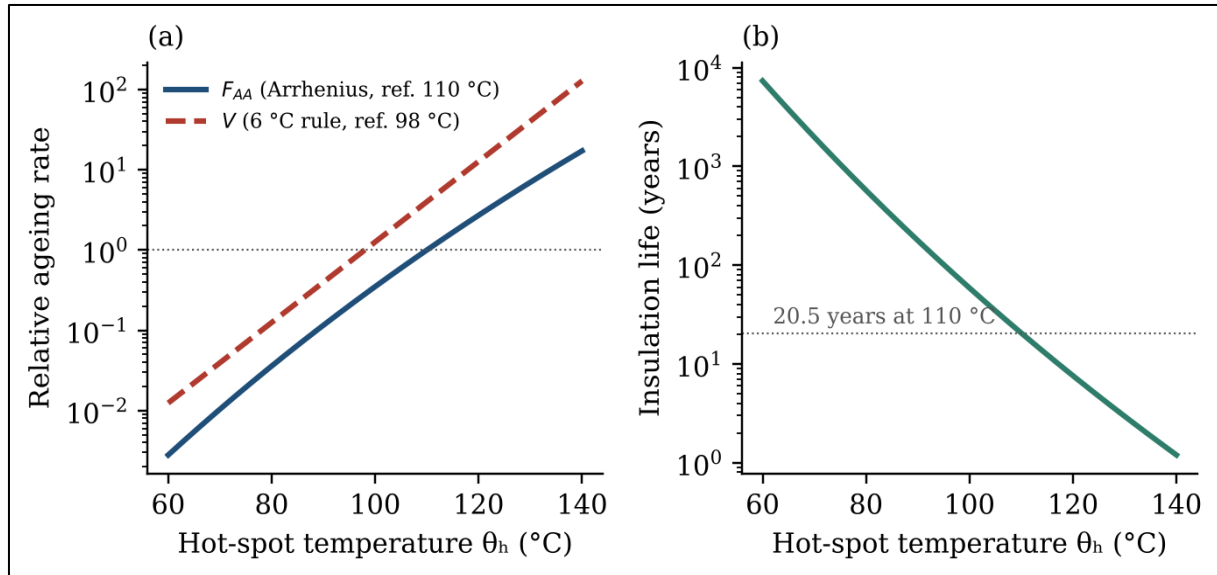


Fig 2 Ageing Rate and Expected Insulation Life as Functions of Hot-Spot Temperature

- Note. Panel (a): Equation 4. Panel (b): Equation 5 with $L_N = 180,000$ h.

III. MATHEMATICAL FORMULATION OF THE CLASSIFIERS

➤ Feature Space

Each oil sample is a vector of gas concentrations with a class label among K states:

$$\mathbf{x} = [\text{H}_2, \text{CO}, \text{C}_2\text{H}_2, \text{C}_2\text{H}_4]^T \in \mathbb{R}^4, \quad y \in \{1, \dots, K\}, \quad z_j = \frac{x_j - \mu_j}{\sigma_j} \quad (6)$$

Standardisation with the training mean μ_j and standard deviation σ_j matters for the distance-based and margin-based models below, because gas concentrations span different ranges. A training set is $D = \{(\mathbf{z}_i, y_i)\}_{i=1}^n$.

➤ Multinomial Logistic Regression

Logistic regression models class probabilities as a softmax of linear scores and estimates the weights by minimising the cross-entropy:

$$P(y = k | \mathbf{z}) = \frac{\exp(\mathbf{w}_k^T \mathbf{z} + b_k)}{\sum_{l=1}^K \exp(\mathbf{w}_l^T \mathbf{z} + b_l)}, \quad \mathcal{L} = -\frac{1}{n} \sum_{i=1}^n \ln P(y = y_i | \mathbf{z}_i) \quad (7)$$

The boundary between two classes is the hyperplane $(\mathbf{w}_k - \mathbf{w}_l)^T \mathbf{z} + b_k - b_l = 0$. Since Equation 1 makes gas ratios, not sums, the discriminating quantity, a linear boundary in concentration space is a strong restriction.

➤ k-Nearest Neighbours

The k-nearest-neighbour rule assigns the majority class among the k training points closest to $\mathbf{z}$ in Euclidean distance:

$$d(\mathbf{z}, \mathbf{z}_i) = \sqrt{\sum_{j=1}^4 (z_j - z_{ij})^2}, \quad \hat{y} = \underset{c}{\operatorname{argmax}} \sum_{i \in N_k(\mathbf{z})} \mathbb{1}(y_i = c) \quad (8)$$

It has no training phase and adapts to any boundary shape, at the cost of sensitivity to scaling and to local label noise.

➤ *Support Vector Machine*

For two classes coded $y_i = \pm 1$, the soft-margin support vector machine solves

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \quad \text{subject to} \quad y_i(w^\top \phi(z_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad (9)$$

Where C penalises margin violations and ϕ maps the input to a feature space. The dual solution depends on the data only through a kernel $\kappa(\mathbf{z}, \mathbf{z}') = \phi(\mathbf{z})^\top \phi(\mathbf{z}')$:

$$f(\mathbf{z}) = \text{sign}(\sum_{i=1}^n \alpha_i y_i \kappa(\mathbf{z}_i, \mathbf{z}) + b), \quad \kappa(\mathbf{z}, \mathbf{z}') = \exp(-\gamma \|\mathbf{z} - \mathbf{z}'\|^2) \quad (10)$$

With the linear kernel the boundary is a hyperplane, as in Equation 7; the radial kernel shown gives curved boundaries. Multi-class problems are handled by one-versus-rest, training K binary machines and choosing the class with the largest score.

➤ *Random Forest*

A decision tree splits the feature space by thresholds on single gases, choosing at each node the split that most reduces the Gini impurity:

$$G(S) = 1 - \sum_{k=1}^K p_k^2, \quad \Delta G = G(S) - \frac{|S_L|}{|S|} G(S_L) - \frac{|S_R|}{|S|} G(S_R) \quad (11)$$

Where p_k is the proportion of class k in node S . Thresholds on individual gases reproduce the logic of key-gas rules, but learned from data. A random forest trains B trees h_b on bootstrap samples with a random subset of features at each split, and aggregates them by vote (Mienye & Sun, 2022):

$$\begin{aligned} \hat{P}(y = c | \mathbf{z}) &= \frac{1}{B} \sum_{b=1}^B \mathbb{1}(h_b(\mathbf{z}) = c), \\ \text{Var}\left(\frac{1}{B} \sum_{b=1}^B h_b\right) &= \rho \sigma^2 + \frac{1-\rho}{B} \sigma^2 \end{aligned} \quad (12)$$

The second relation, for trees of variance σ^2 and pairwise correlation ρ , explains why forests generalise better than single trees and why the number of trees matters (Figure 3). With $B = 4$ the ensemble keeps 25%, 47.5% and 70% of the single-tree variance for $\rho = 0, 0.3$ and 0.6 ; with $B = 100$ these fall to 1%, 30.7% and 60.4%. Beyond a few hundred trees the floor $\rho \sigma^2$ dominates, so decorrelating the trees is worth more than adding trees.

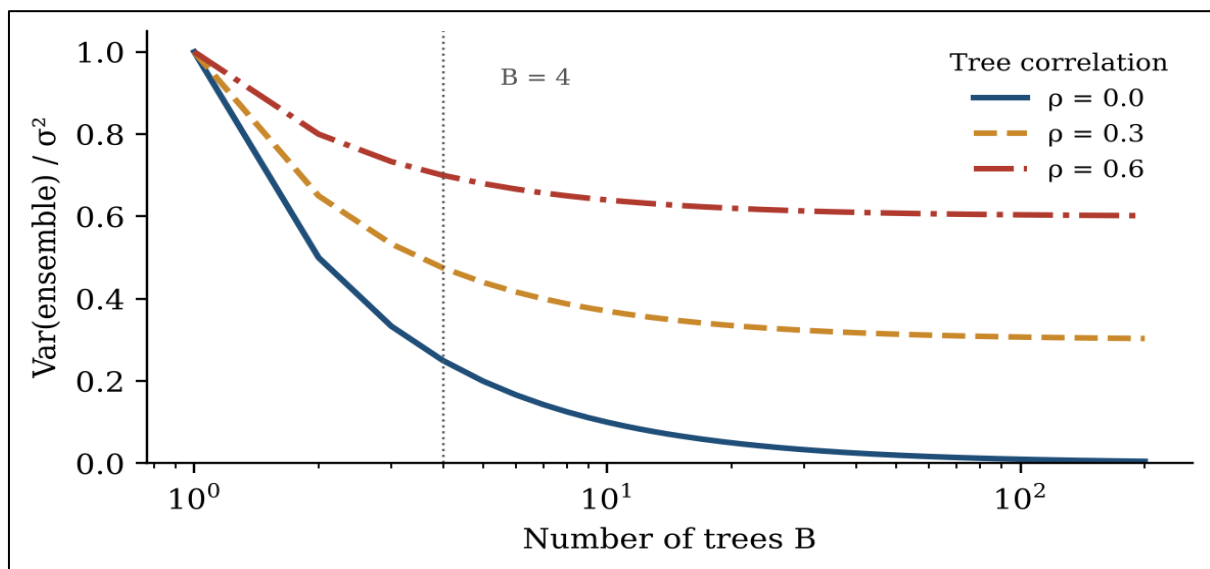


Fig 3 Variance of a Tree Ensemble Relative to a Single Tree, as a Function of the Number of Trees

- *Note.* Equation 12, second relation. The dotted line marks $B = 4$.

➤ *Stacked Generalisation*

Stacking trains M base classifiers $h_1, \dots, h_M$ and a meta-classifier g on their outputs:

$$\mathbf{u}_i = [h_1(\mathbf{z}_i), \dots, h_M(\mathbf{z}_i)], \quad \hat{y}_i = g(\mathbf{u}_i) \quad (13)$$

To avoid the meta-classifier learning from predictions made on data the base models have already seen, $\mathbf{u}_i$ must be produced by cross-validation: each $h_m(\mathbf{z}_i)$ comes from a model fitted without sample i . Stacking can only gain if the base models make different errors (Mienye & Sun, 2022). Figure 4 illustrates the four families on a synthetic two-feature problem: the linear model draws straight boundaries, the others bend them where classes overlap.

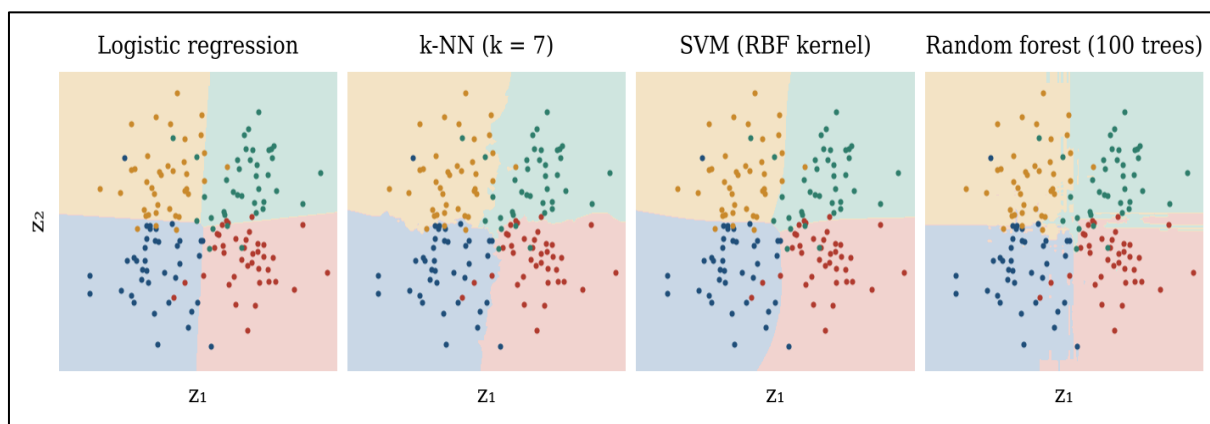


Fig 4 Decision Regions of Four Classifiers on a Synthetic Four-Class Problem

- *Note.* Synthetic standardised data generated for illustration (four Gaussian classes, 150 points each); these are not transformer measurements.

➤ Evaluation

From the confusion matrix **C** (C_{ij} = samples of true class i predicted as j , N = total), accuracy, per-class precision, recall and F1-score are:

$$p_o = \frac{\sum_i C_{ii}}{N}, \quad P_j = \frac{C_{jj}}{\sum_i C_{ij}}, \quad R_i = \frac{C_{ii}}{\sum_j C_{ij}}, \quad F_1 = \frac{2PR}{P+R} \quad (14)$$

Accuracy estimated on N test samples carries sampling uncertainty, quantified by the Wilson interval at 95% ($z = 1.96$):

$$p_o + \frac{z^2}{2N} \pm z \sqrt{\frac{p_o(1-p_o)}{N} + \frac{z^2}{4N^2}} \quad (15)$$

For $N = 1,000$ and p_o near 0.9 the half-width is about 1.9 percentage points, so differences smaller than this between two models are not evidence that one is better.

➤ Synthesis

Table 1 relates each physical mechanism to the gas that reveals it and to the model property needed to exploit it.

Table 1 Correspondence Between Physical Mechanisms, Indicator Gases and Model Requirements

Mechanism	Indicator	Physical law (equation)	Consequence for the model
Low-energy discharge	H ₂	Bond energy, Arrhenius (1)	H ₂ alone is ambiguous: also produced by arcs
Thermal fault, above 500 °C	C ₂ H ₄	Arrhenius (1)	Class depends on ratios: nonlinear boundary
Arc, 800–1,200 °C	C ₂ H ₂	Arrhenius (1)	Shares H ₂ with partial discharge
Paper ageing, above 105 °C	CO	Arrhenius, loss of life (4, 5)	Slow drift: trend matters more than level
Sensor in gas phase	all	Ostwald partition (2)	Convert before comparison

- *Note.* Temperatures are the indicative thresholds given in IEC 60599 (IEC, 2022).

Three points follow. First, because fault classes are separated by ratios of gases whose formation rates differ by orders of magnitude (Figure 1), linear models in raw concentrations are structurally limited, and published comparisons agree that nonlinear and ensemble models perform better (Rao et al., 2021); using gas ratios as inputs is an alternative that also improves robustness to measurement noise (Taha et al., 2021). Second, partial discharge and arcing are both electrical faults that release hydrogen; they are distinguished mainly by acetylene, so a model given few gases should be expected to confuse them. Third, diagnosis and life estimation rest on different physics. Classification uses the instantaneous gas pattern, whereas remaining life is an integral of temperature history (Equation 5) and cannot be obtained from a single gas reading without a kinetic model linking CO production to cellulose degradation.

IV. CONCLUSION

This article has connected the thermochemistry of gas formation, gas–oil partition and cellulose ageing to the definitions of four classifiers, stacking and their evaluation. The numerical illustrations show the sensitivity of insulation life to hot-spot temperature (a factor of 2.71 for 10 K above the reference) and the limited variance reduction offered by very small ensembles. The gas-formation curves rest on illustrative activation energies and the decision regions on synthetic data; they explain the design and do not characterise a particular transformer. The companion article evaluates the models on a four-gas data set and discusses their integration in a substation monitoring system.

REFERENCES

- [1]. Achouch, M., Dimitrova, M., Ziane, K., Sattarpanah Karganroudi, S., Dhoub, R., Ibrahim, H., & Adda, M. (2022). On predictive maintenance in Industry 4.0: Overview, models, and challenges. *Applied*

- Sciences, 12(16), Article 8081. <https://doi.org/10.3390/app12168081>
- [2]. Acikgoz, N., Sazli, M. H., & Kosalay, I. (2025). A comprehensive review of transformer fault diagnosis studies based on dissolved gas analysis: Classical methods, historical development of the devices used, artificial intelligence based methods, accuracy of classifications of predictions. *Elektronika ir Elektrotechnika*, 31(3), 4–21. <https://doi.org/10.5755/j02.eie.39824>
- [3]. Behjat, V., Emadifar, R., Pourhossein, M., Rao, U. M., Fofana, I., & Najjar, R. (2021). Improved monitoring and diagnosis of transformer solid insulation using pertinent chemical indicators. *Energies*, 14(13), Article 3977. <https://doi.org/10.3390/en14133977>
- [4]. Bilyaz, S., Bhati, A., Hamalian, M., Maynor, K., Soori, T., Gattozzi, A., Penney, C., Weeks, D., Xu, Y., Hu, L., Zhu, J. Y., Nelson, J. K., Hebner, R., & Bahadur, V. (2024). Modeling the impact of high thermal conductivity paper on the performance and life of power transformers. *Heliyon*, 10(6), Article e27783. <https://doi.org/10.1016/j.heliyon.2024.e27783>
- [5]. Duanaputri, R., Prasajo, R. A., Wiranti, D., Wiwaha, S. S., Akbar, D. N., & Handani, G. P. C. (2023). Analysis and experiments of insulation loss of transformer kraft paper based on hotspot variations. *International Journal of Electrical Engineering and Applied Sciences*, 6(1). <https://doi.org/10.54554/ijeeas.2023.6.01.004>
- [6]. International Electrotechnical Commission. (2022). *Mineral oil-filled electrical equipment in service – Guidance on the interpretation of dissolved and free gases analysis* (IEC 60599:2022, 4th ed.).
- [7]. Leu, B., Seritan, G., & Enache, B. (2023). A new hybrid model to evaluate the loss of life of power transformers. *Revue Roumaine des Sciences Techniques, Série Électrotechnique et Énergétique*, 68(3), 277–282. <https://doi.org/10.59277/RRST-EE.2023.3.5>
- [8]. Li, Y. (2023). *The state of the art in transformer fault diagnosis with artificial intelligence and dissolved gas analysis: A review of the literature*. arXiv. <https://doi.org/10.48550/arXiv.2304.11880>
- [9]. Mienye, I. D., & Sun, Y. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access*, 10, 99129–99149. <https://doi.org/10.1109/ACCESS.2022.3207287>
- [10]. Nanfak, A., Eke, S., Fofana, I., Meghnefi, F., Ngaleu, G. M., & Kom, C. H. (2024). Traditional fault diagnosis methods for mineral oil-immersed power transformer based on dissolved gas analysis: Past, present and future. *IET Nanodielectrics*. Advance online publication. <https://doi.org/10.1049/nde2.12082>
- [11]. Rao, U. M., Fofana, I., Rajesh, K. N. V. P. S., & Picher, P. (2021). Identification and application of machine learning algorithms for transformer dissolved gas analysis. *IEEE Transactions on Dielectrics and Electrical Insulation*, 28(5), 1828–1835. <https://doi.org/10.1109/TDEI.2021.009770>
- [12]. Taha, I. B. M., Ibrahim, S., & Mansour, D.-E. A. (2021). Power transformer fault diagnosis based on DGA using a convolutional neural network with noise in measurements. *IEEE Access*, 9, 111162–111170. <https://doi.org/10.1109/ACCESS.2021.3102415>
- [13]. Zhang, Y., Tang, Y., Liu, Y., & Liang, Z. (2022). Fault diagnosis of transformer using artificial intelligence: A review. *Frontiers in Energy Research*, 10, Article 1006474. <https://doi.org/10.3389/fenrg.2022.1006474>