Breakdown Voltage Prediction of Transformer Oil under Copper/Iron Contamination Using SVR and KNN Regression: A Baseline Study on Small Experimental Data

Giovanni Dimas Prenata[a], Ahmad Ridho'i[b]

[a,b] *Electrical Engineering, Faculty of Intelligent Electrical and Informatics Technology, Universitas 17 Agustus 1945 Surabaya, Indonesia*

ARTICLE INFO

Article history:
Received 3 June 2026
Received in revised form June 30, 2026
Accepted 6 July 2026
Available online 10 July 2026

Keywords:
Transformer oil
Breakdown voltage
Support vector regression
K-nearest neighbor
Metallic contamination
Small experimental data

ABSTRACT

This study evaluates Support Vector Regression (SVR) and K-Nearest Neighbor (KNN) regression as transparent baseline models for predicting transformer-oil breakdown voltage (BDV) under copper and iron contamination. The dataset contains 324 observations from three oil types, two metallic contaminants, three concentration levels, and 18 repeated SampleIDs. Oil and metal categories were represented by one-hot encoding, while concentration was normalized using training data only. To prevent leakage and examine split sensitivity, model selection and evaluation used nested grouped validation: six outer folds held out three SampleIDs each, and five inner grouped folds selected K for KNN and γ–σ for LS-SVR. SVR achieved $R^2 = 0.016 \pm 0.130$, RMSE = 1.561 ± 0.176 kV, and MAE = 1.254 ± 0.157 kV. KNN produced $R^2 = 0.005 \pm 0.128$, RMSE = 1.569 ± 0.172 kV, and MAE = 1.262 ± 0.161 kV. The SVR reductions in RMSE and MAE were only 0.56% and 0.61%, and the paired per-SampleID error difference was not significant (Wilcoxon $p = 0.640$). Both models reached approximately 95.1% accuracy within ±3 kV, while KNN required substantially lower computation time. The low R^2 values indicate that oil type, metal type, and concentration alone cannot explain the large within-condition BDV variability. Therefore, the models are positioned as preliminary screening and benchmarking tools rather than replacements for standardized BDV testing.

1 Introduction

Transformer oil plays a critical role as both an insulating and cooling medium in oil-filled power transformers. Its dielectric reliability directly affects the operational safety of high-voltage equipment because insulation failure can initiate discharge, damage internal components, and interrupt power delivery. Breakdown voltage (BDV) is a direct indicator of the electric stress that the liquid insulation can withstand before a conductive path is formed. Metallic particles, particularly copper (Cu) from windings and iron (Fe) from cores, tanks, or structural components, can polarize, migrate, distort the local electric field, and facilitate particle bridging and streamer initiation [1]–[3]. Consequently, estimating BDV under controlled metallic contamination is relevant to transformer-oil condition assessment, although standardized acceptance decisions must remain based on laboratory testing such as IEC 60156 [4].

Machine-learning methods are increasingly used in transformer diagnostics, health-index estimation, dissolved-gas interpretation, and dielectric-property assessment [5]–[11]. Regression analysis has been applied to examine the relationship between water content and transformer-oil electrical properties, including breakdown voltage, resistivity, and dielectric dissipation factor [5]. Adaptive and deep-learning approaches have also been used to evaluate dissolved-gas measurements for power-transformer condition assessment [6]. Random-forest models have been applied to transformer-oil quality assessment [7], while combined predictive-maintenance and health-index frameworks have incorporated BDV, moisture, acidity, dissolved gases, and other condition indicators [8]–[11]. These studies demonstrate the value of data-driven methods for transformer condition assessment, although they generally require additional physical, chemical, gas, moisture, or operational measurements that are not always available in limited laboratory datasets.

SVR and KNN were selected in the present study because they provide complementary and interpretable baselines for a small experimental dataset. SVR uses regularization and a kernel function to construct a smooth global nonlinear mapping and is well established for limited-sample regression [12] [13]. KNN is a non-parametric local estimator that predicts from neighboring observations without estimating a large set of model weights [14]. In contrast, ensemble methods such as Random Forest and XGBoost require multiple trees, Gaussian Process Regression requires covariance-model assumptions and matrix operations, and multilayer neural networks introduce additional architecture and optimization choices. The purpose of this work is therefore not to claim that SVR and KNN are universally superior, but to establish whether a global kernel model provides a meaningful advantage over a simple local-similarity model under the same controlled dielectric dataset and validation protocol.

Previous studies by the authors examined the same physical domain from different analytical perspectives. A DFFNN study compared gradient-based, Genetic Algorithm, and Grey Wolf Optimizer training for BDV prediction under Cu/Fe contamination [15], whereas a clustering study used K-Means and Fuzzy C-Means to identify hard and transitional dielectric regimes [16]. The present study is distinct because it evaluates two supervised regression baselines, corrects nominal-feature representation through one-hot encoding, and assesses generalization using nested grouped validation by SampleID rather than a single deterministic train–test split.

The remaining gap is a reproducible baseline comparison across multiple oil types, both Cu and Fe contaminants, and three concentration levels using one unified experimental dataset. Earlier transformer studies frequently emphasized fault classification, health-index construction, spectroscopic prediction, or individual contamination scenarios [5]–[11]. They provide limited evidence on whether a transparent distance-based regressor can match a kernel-based regressor when repeated laboratory measurements from the same SampleID are kept together during validation.

Accordingly, this study contributes: (1) a leakage-resistant nested grouped-validation design in which each SampleID appears in an outer test fold exactly once; (2) one-hot

Corresponding Author: gprenata@untag-sby.ac.id

Doi : 10.30996/elsains.v8i1.133785

representation of nominal oil and metal categories; (3) sensitivity analysis across six alternative grouped splits; (4) repeated-measures statistical analysis of oil, metal, and concentration effects; and (5) comparison of predictive error, tolerance-based accuracy, computational cost, and interpretability. The intended engineering role is preliminary screening and decision support: the estimates may help prioritize confirmatory tests or identify conditions requiring closer inspection, but they are not proposed as replacements for standardized BDV measurement.

2 Literature Study

2.1. Transformer Oil Breakdown Voltage and Metallic Contamination

Transformer oil functions as both an electrical insulating medium and a heat-transfer medium in oil-filled power transformers. Its dielectric quality is commonly evaluated using Breakdown Voltage (BDV), which represents the voltage at which the oil loses its insulating capability and a conductive discharge path is formed between the electrodes. Under an approximately uniform electric field, the average electric-field intensity can be expressed as :

$$E_0 = \frac{V}{d} \quad (1)$$

where E_0 is the average electric-field intensity, V is the applied voltage, and d is the electrode gap. Electrical breakdown occurs when the local electric field reaches or exceeds the critical dielectric strength of the oil [1].

Metallic particles such as copper and iron can intensify the local electric field around their surfaces. This phenomenon can be represented conceptually by :

$$E_{local} = \beta_p E_0 \quad (2)$$

where β_p is the particle-induced field-enhancement factor. For uncontaminated oil, β_p is relatively close to unity, whereas suspended conductive particles may increase its value. Consequently, the approximate breakdown voltage can be expressed as :

$$V_{BDV} \approx \frac{E_{crit} d}{\beta_p} \quad (3)$$

where E_{crit} denotes the critical electric-field strength of the oil. Equation (3) indicates that an increase in the field-enhancement factor caused by metallic contamination may reduce the measured BDV [2]. In actual transformer oil, this relationship is influenced by particle concentration, particle geometry, moisture, oil type, temperature, electrode condition, and microscopic discharge-channel formation [1] [2] [3]. Therefore, BDV behavior generally exhibits stochastic and nonlinear characteristics rather than following a single deterministic physical equation.

Copper particles may originate from transformer windings, whereas iron particles may be released from the magnetic core, tank, or other internal components. Under an applied electric field, these particles become polarized, migrate within the oil, and may form temporary particle bridges between electrodes. These mechanisms distort the field distribution and facilitate partial-discharge initiation, thereby increasing the probability of breakdown [2]. Previous studies have mainly examined these effects through experimental dielectric testing, while direct regression-based prediction across multiple oil types and contamination conditions remains limited.

2.2. Support Vector Regression

Support Vector Regression is a supervised learning technique for estimating continuous outputs [12]. In this study, SVR is used to model the nonlinear relationship between oil type, contaminant metal, contaminant concentration, and measured BDV. The general SVR regression function is written as:

$$f(x) = w^T \emptyset(x) + b$$

where x is the input vector, ϕ(x) maps the original input into a higher-dimensional feature space, w is the weight vector, and b is the bias.

Because the implementation uses Least Squares Support Vector Regression, its optimization problem can be expressed as [13]:

$$\min_{w,b,\{e_i\}} \left[\frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^{n} e_i^2\right] \quad (4)$$

subject to

$$y_i = w^T \emptyset(x_i) + b + e_i, \quad i = 1,2,3,\ldots\ldots,n \quad (5)$$

where y_i is the measured BDV, e_i is the prediction error, n is the number of training observations, and γ is the regularization parameter. The first term in (4) controls model complexity, whereas the second term penalizes squared prediction errors. A large γ emphasizes fitting accuracy, while a smaller value produces a smoother model with stronger regularization.

The nonlinear mapping is implemented using the Radial Basis Function kernel:

$$K(x_i, x_j) = exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (6)$$

where σ is the kernel-width parameter. A small σ makes the model highly sensitive to local variations, whereas a large σ produces a smoother regression surface. After solving the optimization problem, the SVR prediction can be written as:

$$\hat{y}(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) + b \quad (7)$$

where α_i represents the contribution of the i-th training observation.

The parameters γ and σ strongly influence prediction performance. In this study, they are selected through five inner grouped folds within each outer grouped-validation fold. The held-out outer test SampleIDs are not used during hyperparameter selection. This nested procedure minimizes inner-validation error while preserving an independent outer test fold. SVR represents a global nonlinear mapping approach because it constructs a regression function covering the entire input space.

2.3. K-Nearest Neighbor Regression

K-Nearest Neighbor Regression is a non-parametric and instance-based prediction method. Unlike SVR, KNN does not construct an explicit regression function during training. Instead, it stores the training observations and estimates the output of a new sample from the outputs of nearby observations [14].

For an input vector x and i-th training observation x_i, the Euclidean distance is calculated as

$$d(x, x_i) = \sqrt{\sum_{j=1}^{p} (x_j - x_{ij})^2} \quad (8)$$

where p is the number of input features, x_j is the j-th feature of the test sample, and x_{ij} is the corresponding feature of the i-th training sample.

After calculating all distances, the K observations with the smallest distances are selected. The predicted BDV is then obtained using the arithmetic mean:

$$\hat{y}(x) = \frac{1}{K} \sum_{i \in N_K(x)} y_i \quad (9)$$

where $N_K(x)$ denotes the set of K nearest training observations and y_i is the measured BDV associated with the i-th neighbor.

The number of neighbors affects the bias–variance characteristics of the prediction. A very small K may cause the model to become sensitive to individual measurements or

Corresponding Author: gprenata@untag-sby.ac.id
Doi: 10.30996/elsains.v8i1.133785

experimental noise, while a very large K may excessively smooth the prediction and ignore local differences. Therefore, the optimal number of neighbors is selected by:

$$K^* = \underset{K \in \{1,2,3,\ldots.45\}}{\arg min} RMSE_{inner}(K) \quad (10)$$

with

$$RMSE_{inner}(K) = \sqrt{\frac{1}{n_v}\sum_{i=1}^{n_v}(y_i - \hat{y}_i^{(K)})^2} \quad (11)$$

where n_v is the number of inner-validation observations. The outer test fold is used only after K has been selected. KNN therefore represents a local similarity-based approach, because its predictions depend directly on neighboring experimental conditions rather than on a global mathematical function.

2.4. Comparison of Kernel-Based and Distance-Based Regression

The primary theoretical difference between SVR and KNN Regression lies in how the relationship between input variables and BDV is represented. SVR constructs a global nonlinear regression function using an RBF kernel:

$$\hat{y}SVR = f_{global}(x) \quad (12)$$

whereas KNN estimates BDV from a local neighborhood:

$$\hat{y}KNN = f_{local}(x|N_K(x)) \quad (13)$$

SVR is theoretically advantageous when the relationship between contamination parameters and BDV follows a continuous nonlinear pattern. Conversely, KNN may be more appropriate when the experimental observations form local clusters in which similar oil, metal, and concentration conditions produce similar BDV values.

Previous transformer-oil studies have mainly focused on experimental characterization, statistical failure analysis, and classification of oil condition monitoring [2], [3]. Recent transformer studies have employed regression, artificial neural networks, support-vector methods, random forests, and deep-learning techniques for transformer-oil assessment, dissolved-gas interpretation, health-index estimation, and predictive maintenance [5]–[11]. However, these studies commonly use richer physical, chemical, spectroscopic, or operational variables. The present work therefore focuses on a controlled baseline comparison using only oil type, metal type, and contaminant concentration. The comparative behavior of these two regression paradigms for BDV prediction under controlled copper and iron contamination has not been sufficiently examined.

Accordingly, the present study compares SVR and KNN Regression using identical input features, training and testing samples, and evaluation metrics. This comparison is intended to determine whether transformer-oil BDV under metallic contamination is better represented by a global nonlinear kernel function or by local similarity among experimental observations. The contribution of the study is therefore not the proposal of a new algorithm, but the identification of the most appropriate baseline regression characteristic for a limited experimental dielectric dataset.

3 Methodology

3.1. K-Nearest Neighbor Regression

K-Nearest Neighbor (KNN) regression is used as a transparent local baseline for predicting BDV. Each observation contains oil type, contaminant metal, concentration, and the measured BDV target. Because oil and metal are nominal variables, they are represented by one-hot vectors rather than ordinal integer codes. The six-dimensional input is [PSP, Pri Apar, Shell Diala, Cu, Fe, normalized concentration]. Concentration is Min–Max scaled using parameters fitted only on the corresponding training subset.

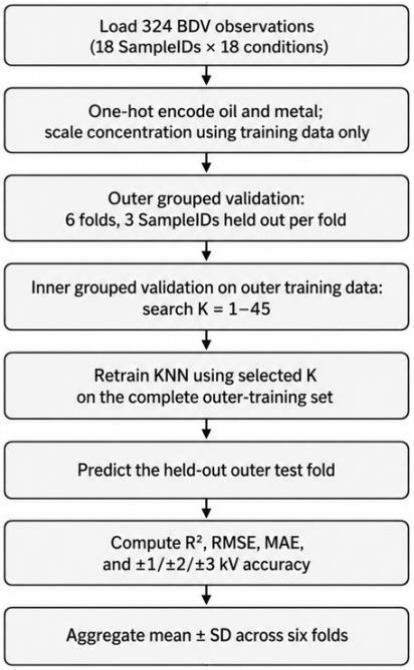

Figure 1. Nested grouped-validation workflow for KNN regression

The dataset contains 324 rows formed by 18 SampleIDs measured under 18 combinations of three oils, two metals, and three concentration levels. All 18 rows belonging to one SampleID are treated as a group. This grouping prevents repeated observations from the same SampleID from being divided between training and testing data.

Sensitivity to the original split is assessed through six outer grouped folds. Each fold holds out three SampleIDs (54 rows) for testing and uses the remaining 15 SampleIDs (270 rows) for training. The test groups are IDs 1–3, 4–6, 7–9, 10–12, 13–15, and 16–18, so every SampleID appears in an outer test fold exactly once. Hyperparameter tuning is performed only within the outer-training data using five inner grouped folds.

For each test observation, Euclidean distance is computed in the six-dimensional encoded feature space. Training observations are sorted using distance and a deterministic row index; the target value is never used to break equal-distance ties. This avoids the target-dependent ordering present in the original implementation.

The predicted BDV is the arithmetic mean of the selected neighbors. Candidate values K = 1,…,45 are evaluated using the pooled RMSE from the five inner grouped folds. The K with the lowest inner-validation RMSE is then used to retrain KNN on the complete outer-training set before prediction of the untouched outer test fold.

Across the six outer folds, the selected K values were 22, 23, 24, 25, 24, and 24, giving 23.67 ± 1.03. Because the optimum no longer occurs at the search boundary, the revised analysis confirms that the previously reported K = 15 was influenced by the restricted candidate range.

Outer-fold performance is quantified using R^2, RMSE, MAE, and the percentages of predictions with absolute errors not exceeding 1, 2, and 3 kV. Mean and sample standard deviation are reported across the six outer folds. This nested design ensures that K selection is independent of each outer test fold and provides a direct sensitivity analysis across alternative SampleID holdouts.

Corresponding Author: gprenata@untag-sby.ac.id

Doi : 10.30996/elsains.v8i1.133785

3.2. Support Vector Regression

Least Squares Support Vector Regression with an RBF kernel is used as the global nonlinear baseline. It uses the same six-dimensional one-hot and normalized feature representation and the same six outer grouped folds as KNN, ensuring that model comparisons are not confounded by different data partitions or preprocessing.

Within each outer-training set, five inner grouped folds select the regularization and kernel-width parameters. All preprocessing parameters are fitted on the inner-training subset and applied unchanged to its validation subset. After parameter selection, the model is retrained on all 270 rows of the corresponding outer-training fold and evaluated on the 54 held-out rows.

Oil and metal indicators already have values of zero or one, while concentration is scaled to [0,1] using the training-only minimum and maximum. This representation removes the artificial ordinal distances produced by codes such as PSP = 0, Pri Apar = 1, and Shell Diala = 2.

The LS-SVR model maps the encoded inputs through the RBF kernel. The final expanded grid is γ = {0.01, 0.05, 0.1, 0.5, 1, 10, 100, 1000} and σ = {0.1, 0.2, 0.5, 1, 2, 5}. The expanded lower γ boundary was introduced after the initial reviewer-oriented run selected γ = 0.1 in two folds. The rerun retained exactly the same selected parameters and metrics, confirming that the optimum was not caused by a restricted grid.

The parameter pair minimizing pooled inner grouped-validation RMSE is selected independently for each outer fold. Folds 1–4 selected γ = 1 and σ = 1, while Folds 5–6 selected γ = 0.1 and σ = 0.5. None of the selected values lies on the boundary of the expanded grid.

The final LS-SVR system is solved using a Cholesky-based linear-system procedure. Predictions and all evaluation metrics are recorded separately for each outer fold, followed by mean and sample standard deviation across folds.

Using identical grouped folds permits a paired comparison of SVR and KNN under the same held-out SampleIDs.

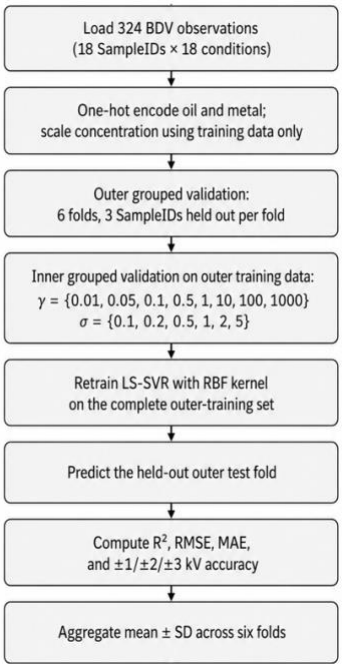

Figure 2. Nested grouped-validation workflow for LS-SVR

3.3. Statistical Analysis and Model Comparison

Because every SampleID contributes measurements to all 18 oil–metal–concentration conditions, a three-factor repeated-measures ANOVA is used with within-subject factors oil type (3 levels), metal type (2 levels), and concentration (3 levels). Significant interactions are followed by paired comparisons of Cu and Fe within each oil after averaging the three concentration measurements for each SampleID. Holm correction is applied to the three paired comparisons, and 95% confidence intervals are reported.

For model comparison, the absolute errors are first averaged within each SampleID, producing 18 paired MAE values for SVR and KNN. A two-sided Wilcoxon signed-rank test evaluates whether the models differ systematically. Statistical significance is assessed at $p < 0.05$. Runtime is reported only for the present C++ implementation and hardware environment and is not interpreted as a universal algorithmic benchmark.

4 Discussion

The revised analysis evaluates all 324 observations through six outer grouped folds rather than relying only on SampleIDs 16–18. Each outer fold contains 270 training rows and 54 unseen test rows, and all measurements belonging to a SampleID remain together. This design directly tests whether the results are sensitive to the original deterministic split while preserving the repeated-measurement structure.

The dataset comprises 18 experimental conditions and 18 repeated SampleIDs per condition. Table 1 shows mean BDV values from 15.20 kV for Shell Diala–Cu at 0.20 g to 17.62 kV for PSP–Fe at 0.10 g, with standard deviations from 0.99 to 1.94 kV. The comparatively large dispersion within nominally identical conditions is important because the regression models receive the same three physical descriptors for repeated measurements that nevertheless have different BDV targets.

Figure 3 summarizes Cu and Fe measurements for each oil after averaging the three concentrations within each SampleID. Error bars show 95% confidence intervals across the 18 SampleIDs. The plot is descriptive; formal inference is based on the repeated-measures analysis reported below.

Table 1. Summary of BDV measurements for each experimental condition (n = 18 per condition)

Condition	n	Mean (kV)	Std (kV)	Min (kV)	Max (kV)
PSP_Cu_0.10g	18	15.34	1.32	12.8	16.9
PSP_Cu_0.15g	18	16.05	1.90	12.9	19.9
PSP_Cu_0.20g	18	17.14	1.94	13.7	20.7
PSP_Fe_0.10g	18	17.62	1.45	14.1	19.6
PSP_Fe_0.15g	18	17.26	1.17	15.6	18.9
PSP_Fe_0.20g	18	16.49	1.10	14.7	18.8
Pri_Apar_Cu_0.10g	18	16.77	1.54	14.1	19.3
Pri_Apar_Cu_0.15g	18	16.86	1.43	14.8	19.3
Pri_Apar_Cu_0.20g	18	16.99	1.89	13.3	20.7
Pri_Apar_Fe_0.10g	18	15.80	1.32	13.2	18.2
Pri_Apar_Fe_0.15g	18	15.52	1.02	13.9	18.1
Pri_Apar_Fe_0.20g	18	15.34	1.32	12.8	16.9
Shell_Diala_Cu_0.10g	18	16.34	1.67	13.7	19.9
Shell_Diala_Cu_0.15g	18	15.35	1.55	12.8	18.7
Shell_Diala_Cu_0.20g	18	15.20	1.82	12.7	18.5
Shell_Diala_Fe_0.10g	18	16.27	1.77	13.6	19.4
Shell_Diala_Fe_0.15g	18	17.09	1.04	15.8	19.4
Shell_Diala_Fe_0.20g	18	17.09	0.99	14.8	18.5

Corresponding Author: gprenata@untag-sby.ac.id
Doi: 10.30996/elsains.v8i1.133785

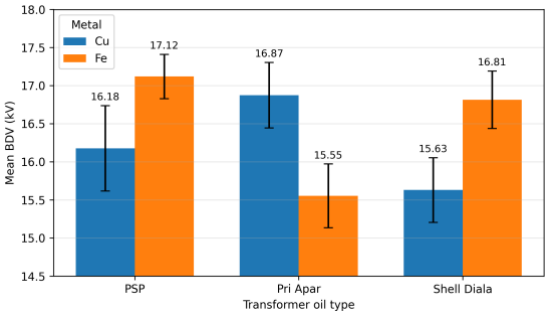

Figure 3. Mean BDV for Cu- and Fe-contaminated oils with 95% confidence intervals

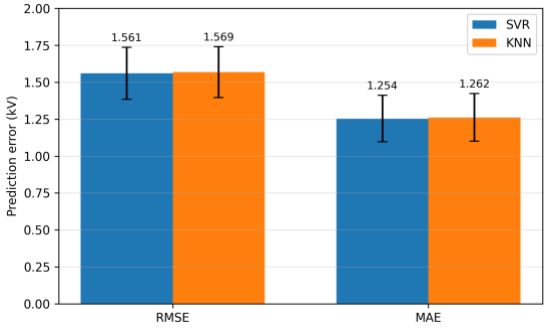

Figure 4. Across-fold RMSE and MAE for SVR and KNN (mean ± SD)

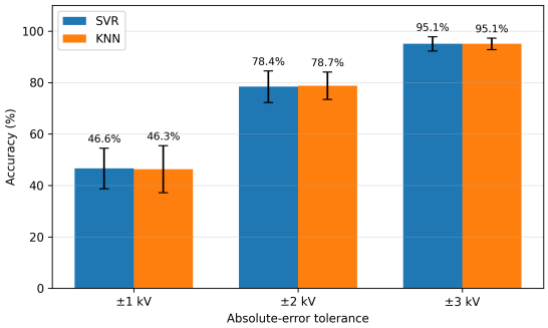

Figure 5. Tolerance-based accuracy of SVR and KNN (mean ± SD)

The repeated-measures ANOVA in Table 2 identifies significant main effects of oil type, $F(2,34) = 4.273$, $p = 0.022$, and metal type, $F(1,17) = 5.155$, $p = 0.036$, but no concentration main effect, $F(2,34) = 0.007$, $p = 0.993$. More importantly, the oil × metal interaction is highly significant, $F(2,34) = 23.889$, $p < 0.001$, and the three-way oil × metal × concentration interaction is also significant, $F(4,68) = 8.327$, $p < 0.001$. Holm-adjusted paired comparisons confirm that Fe exceeds Cu for PSP by 0.943 kV (95% CI: 0.518–1.367 kV, adjusted $p < 0.001$) and for Shell Diala by 1.185 kV (95% CI: 0.549–1.822 kV, adjusted $p = 0.001$), whereas Fe is lower than Cu for Pri Apar by 1.320 kV (95% CI: −1.880 to −0.760 kV, adjusted $p < 0.001$). Thus, the contaminant effect cannot be generalized independently of the base oil and concentration.

The opposite Cu/Fe trends are physically plausible because particle-induced breakdown is controlled not only by metal conductivity but also by oil-dependent viscosity, molecular composition, additive package, moisture affinity, charge mobility, and particle dispersion. Conductive particles can distort the local field and facilitate temporary bridges or streamer growth [1], [2], while the base liquid influences particle motion, charge relaxation, and discharge-channel development [1]–[3].

Table 2. Three-factor repeated-measures ANOVA for BDV

Effect	F	df	p-value
Oil type	4.273	2, 34	0.022
Metal type	5.155	1, 17	0.036
Concentration	0.007	2, 34	0.993
Oil × metal	23.889	2, 34	<0.001
Oil × concentration	0.281	4, 68	0.889
Metal × concentration	1.128	2, 34	0.336
Oil × metal × concentration	8.327	4, 68	<0.001

Table 3. Sensitivity analysis across six grouped outer folds

Fold	Test IDs	SVR R^2	SVR RMSE	KNN R^2	KNN RMSE	Best K
1	1–3	−0.092	1.689	−0.116	1.708	22
2	4–6	−0.109	1.657	−0.071	1.629	23
3	7–9	−0.045	1.482	−0.118	1.532	24
4	10–12	0.016	1.458	0.058	1.426	25
5	13–15	0.235	1.304	0.199	1.333	24
6	16–18	0.092	1.774	0.077	1.789	24

Table 4. Across-fold model performance (mean ± SD)

Metric	SVR	KNN
R^2	0.016 ± 0.130	0.005 ± 0.128
RMSE (kV)	1.561 ± 0.176	1.569 ± 0.172
MAE (kV)	1.254 ± 0.157	1.262 ± 0.161
Accuracy ±1 kV (%)	46.60 ± 7.89	46.30 ± 9.15
Accuracy ±2 kV (%)	78.40 ± 6.16	78.70 ± 5.34
Accuracy ±3 kV (%)	95.06 ± 2.79	95.06 ± 2.24

The significant interactions therefore support an oil-specific interpretation rather than a universal ranking of Cu and Fe.

Table 3 presents the six outer grouped-fold results. SVR RMSE ranges from 1.304 to 1.774 kV, while KNN ranges from 1.333 to 1.789 kV. The original holdout, SampleIDs 16–18, is Fold 6 and yields the largest RMSE for both models, confirming the reviewer's concern that a single split could overstate or understate performance. The six-fold mean and standard deviation provide a more defensible estimate of generalization. KNN selects K values from 22 to 25 rather than the former boundary value of 15, while LS-SVR consistently selects $(\gamma, \sigma) = (1,1)$ in four folds and (0.1,0.5) in two folds.

The across-fold error comparison is visualized in Figure 4 and summarized in Table 4. SVR obtains $R^2 = 0.016 \pm 0.130$, RMSE = 1.561 ± 0.176 kV, and MAE = 1.254 ± 0.157 kV. KNN obtains $R^2 = 0.005 \pm 0.128$, RMSE = 1.569 ± 0.172 kV, and MAE = 1.262 ± 0.161 kV.

SVR reduces RMSE by only 0.56% and MAE by 0.61% relative to KNN. At the SampleID level, the mean paired MAE difference (SVR − KNN) is −0.0077 kV with a 95% confidence interval from −0.0396 to 0.0241 kV; the Wilcoxon signed-rank test

is not significant ($p = 0.640$). Therefore, the numerical advantage of SVR is neither statistically nor practically sufficient to select it on accuracy alone. KNN is easier to interpret through its neighbors and requires 0.286 s per fold on average, whereas nested LS-SVR tuning requires 71.750 s per fold in the present implementation, approximately 251 times longer. This runtime comparison is implementation-specific, but it highlights the lower computational burden of KNN for this small baseline task.

Tolerance-based accuracy provides an engineering-oriented description of absolute prediction deviation. SVR achieves 46.60 ± 7.89%, 78.40 ± 6.16%, and 95.06 ± 2.79% within ±1, ±2, and ±3 kV, respectively. KNN achieves 46.30 ± 9.15%, 78.70 ± 5.34%, and 95.06 ± 2.24%. These nearly identical values reinforce the conclusion that the models are comparable.

Figure 5 compares tolerance-based accuracy for both models across the six folds. Approximately 95% of predictions fall within ±3 kV, but only about 46% fall within ±1 kV. Accordingly, these models should be interpreted as coarse preliminary estimators. They may support prioritization of laboratory retesting or comparison of contamination scenarios, but they cannot replace IEC 60156 testing or independently determine whether an oil sample meets an operational acceptance threshold [16].

The low R^2 values require a more complete interpretation than the original manuscript provided. A decomposition by the 18 nominal conditions shows that only 21.65% of total BDV variation occurs between conditions, while 78.35% occurs among repeated measurements sharing identical oil, metal, and concentration inputs. Because the model receives identical predictors for these repetitions, it cannot explain variation arising from unmeasured factors such as particle size and spatial distribution, moisture, temperature, dissolved gases, electrode surface condition, or stochastic streamer initiation. Negative R^2 values in several folds therefore indicate that the three available predictors are insufficient to explain individual-measurement variance, not that low variance alone guarantees acceptable prediction.

The comparable errors of SVR and KNN indicate that additional kernel complexity does not recover information absent from the feature set. SVR provides a smooth global estimate and slightly lower mean error, whereas KNN offers direct local interpretation and substantially lower computational cost. Neither model should be presented as decisively superior; their value is as complementary baselines that reveal the predictive ceiling imposed by the available descriptors.

The revised KNN search also changes the earlier interpretation of $K = 15$. With candidates extended to 45 and selected only by inner grouped validation, the optimal K is 23.67 ± 1.03. The larger neighborhood smooths stochastic repetition-level variation and is stable across outer folds, but it also confirms that predictions are driven by condition-level averaging rather than detailed individual-sample physics.

From a physical standpoint, the models summarize the combined effects of oil, contaminant, and concentration but do not explicitly represent field enhancement, particle transport, moisture-assisted charge conduction, or streamer propagation. The significant interaction effects and large within-condition variability show that future models should include continuous physical measurements rather than simply increasing algorithmic complexity.

For condition-based maintenance, a practical workflow would use the regression output as an auxiliary screening signal alongside measured BDV, moisture, acidity, dielectric dissipation factor, and DGA or health-index indicators [5]–[10]. A prediction differing substantially from recent laboratory measurements could trigger repeat sampling or closer inspection, while stable predictions could help prioritize limited testing resources. However, deployment for online condition monitoring would require prospective data, time-stamped operating variables, and external validation on transformers not represented in this laboratory dataset.

The main limitations are the small number of SampleIDs, the absence of particle-size, moisture, temperature, and electrode-condition measurements, and the use of only three transformer-oil products. Future work should increase the number of independently prepared samples, record these continuous factors, compare additional regressors under the same nested grouped protocol, and evaluate calibration and uncertainty before maintenance decisions are attempted.

Overall, the revised results support a cautious conclusion: SVR and KNN provide comparable preliminary BDV estimates, while the feature set—not the choice between these two algorithms—is the principal limitation on explanatory performance.

5 Conclusion

This study established a leakage-resistant baseline comparison of LS-SVR and KNN regression for transformer-oil BDV under Cu/Fe contamination. The 324 observations were evaluated using one-hot nominal encoding and nested grouped validation across six outer SampleID folds. SVR achieved $R^2 = 0.016 \pm 0.130$, RMSE = 1.561 ± 0.176 kV, and MAE = 1.254 ± 0.157 kV, while KNN achieved $R^2 = 0.005 \pm 0.128$, RMSE = 1.569 ± 0.172 kV, and MAE = 1.262 ± 0.161 kV. The SVR improvements in RMSE and MAE were only 0.56% and 0.61%, and the paired per-SampleID error difference was not significant ($p = 0.640$). Thus, the two models should be regarded as practically comparable rather than as evidence of a decisive winner.

Both models placed approximately 95.1% of predictions within ±3 kV, but only about 46% within ±1 kV. KNN offered substantially lower computation time and direct neighbor-based interpretation, whereas SVR provided a smooth global estimate. The revised KNN optimum of 23.67 ± 1.03 and the stable expanded-grid SVR parameters confirm that the results are not artifacts of the original search boundaries. Nevertheless, the very low R^2 values and the finding that 78.35% of total BDV variation occurs within nominally identical conditions show that oil type, metal type, and concentration are insufficient for high-precision individual prediction. The proposed models are therefore suitable only as preliminary screening and benchmarking tools, not replacements for standardized BDV testing. Future studies should include moisture, temperature, particle size and distribution, dissolved gases, and electrode condition, together with external transformer-level validation.

Author Contributions

Giovanni Dimas Prenata contributed to the conceptualization, methodology, software development, validation, formal analysis, investigation, data curation, visualization, writing of the original draft, and project administration. Ahmad Ridho'i contributed to resource provision, validation, supervision, and writing—review and editing. All authors have read and approved the final version of the manuscript.

Acknowledgment

The authors would like to express their gratitude to the Faculty of Intelligent Electrical and Informatics Technology, Universitas 17 Agustus 1945 Surabaya, for the academic and laboratory support provided during this study. The authors also thank the laboratory

Corresponding Author: gprenata@untag-sby.ac.id
Doi: 10.30996/elsains.v8i1.133785

personnel and colleagues who assisted with transformer oil breakdown-voltage testing, experimental data preparation, and technical discussions related to this work.

References

[1] B. Cao, J.-W. Dong, and M.-H. Chi, "Electrical breakdown mechanism of transformer oil with water impurity: Molecular dynamics simulations and first-principles calculations," *Crystals*, vol. 11, no. 2, Art. no. 123, 2021, doi: 10.3390/cryst11020123.

[2] I. M. Y. Negara, D. Fahmi, D. A. Asfani, I. G. N. Satriyadi Hernanda, M. Wahyudi, and M. N. Ibrahim, "Effect of floating metallic particles in pre-breakdown and breakdown characteristics of oil transformer under DC voltage," *Energies*, vol. 14, no. 12, Art. no. 3669, 2021, doi: 10.3390/en14123669.

[3] U. Elele, A. Nekahi, A. Arshad, and I. Fofana, "Towards online ageing detection in transformer oil: A review," *Sensors*, vol. 22, no. 20, Art. no. 7923, 2022, doi: 10.3390/s22207923.

[4] International Electrotechnical Commission, Insulating Liquids—Determination of the Breakdown Voltage at Power Frequency—Test Method, *IEC 60156:2018*, Geneva, Switzerland, 2018.

[5] S. Abdi, N. Harid, L. Safiddine, A. Boubakeur, and A. M. Haddad, "The correlation of transformer oil electrical properties with water content using a regression approach," *Energies*, vol. 14, no. 8, Art. no. 2089, 2021, doi: 10.3390/en14082089.

[6] D. A. Barkas, S. D. Kaminaris, K. K. Kalkanis, G. Ch. Ioannidis, and C. S. Psomopoulos, "Condition assessment of power transformers through DGA measurements evaluation using adaptive algorithms and deep learning," *Energies*, vol. 16, no. 1, Art. no. 54, 2023, doi: 10.3390/en16010054.

[7] M. E. A. Senoussaoui, M. Brahami, and I. Fofana, "Transformer oil quality assessment using random forest with feature engineering," *Energies*, vol. 14, no. 7, Art. no. 1809, 2021, doi: 10.3390/en14071809.

[8] M. Badawi et al., "Reliable estimation for health index of transformer oil based on novel combined predictive maintenance techniques," *IEEE Access*, vol. 10, pp. 25954–25972, 2022, doi: 10.1109/ACCESS.2022.3156102.

[9] N. R. Arsya, Suwarno, R. A. Prasojo, G. A. Sudrajad, and A. Abu-Siada, "A holistic integration of conventional and machine learning techniques to enhance the analysis of power transformer health index considering data unavailability," *IEEE Access*, vol. 12, pp. 124549–124561, 2024, doi: 10.1109/ACCESS.2024.3450110.

[10] Suwarno, H. Sutikno, R. A. Prasojo, and A. Abu-Siada, "Machine learning based multi-method interpretation to enhance dissolved gas analysis for power transformer fault diagnosis," *Heliyon*, vol. 10, no. 4, Art. no. e25975, 2024, doi: 10.1016/j.heliyon.2024.e25975.

[11] R. Zemouri, "Power transformer prognostics and health management using machine learning: A review and future directions," *Machines*, vol. 13, no. 2, Art. no. 125, 2025, doi: 10.3390/machines13020125.

[12] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004, doi: 10.1023/B:STCO.0000035301.49549.88.

[13] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, 1999, doi: 10.1023/A:1018628609742.

[14] C. J. Stone, "Consistent nonparametric regression," *The Annals of Statistics*, vol. 5, no. 4, pp. 595–620, 1977, doi: 10.1214/aos/1176343886.

[15] G. D. Prenata, "Predicting breakdown voltage of transformer oil under copper/iron contamination: A comparative study of gradient vs. metaheuristic training," *ELKHA: Jurnal Teknik Elektro*, vol. 18, no. 1, pp. 65–73, Apr. 2026.

[16] G. D. Prenata and A. Ridho'i, "Evaluasi komparatif K-Means dan Fuzzy C-Means sebagai metode klasterisasi hard dan soft untuk analisis perilaku dielektrik minyak transformator terkontaminasi logam," *JuTEkS*, vol. 13, no. 1, pp. 422–432, 2026.

Corresponding Author: gprenata@untag-sby.ac.id

Doi : 10.30996/elsains.v8i1.133785