MODERN PREDICTIVE MAINTENANCE OF POWER TRANSFORMERS: MACHINE-LEARNING FAULT DIAGNOSIS FROM DISSOLVED GAS ANALYSIS, AND HOW FAR IT CAN BE TRUSTED

Preprint, not peer reviewed.

Adu-Gyamfi Kwadwo, Department of Electrical and Electronic Engineering, KNUST, Ghana. November 2026. The study extends the final-year project "Modern predictive maintenance of power transformers" (KNUST, 2024), supervised by Dr. E. K. Anto.

Abstract

Power transformers usually develop faults slowly, and an unplanned failure means an outage and an expensive repair. Predictive maintenance tries to catch the fault early. Dissolved gas analysis (DGA) is the main tool for this, and machine learning is now used to read it. This paper tests how far that can be trusted. A random forest trained on 589 dissolved gas records reached 83.8% accuracy on a held-out test set and beat k-nearest neighbours, a support vector machine and a neural network. The study keeps that result and then asks what it is worth. The same forest, on the same gases, scores anywhere between 55% and 99% depending on how it is tested. It reaches 99.2% when the labels come from a rule, 89.3% when synthetic samples are made before splitting the data, 94.4% on one data source, 83.2% under a clean protocol, and 55.5% on 465 records from other sources. Inside one utility, a random split of records gives 93.2%, while keeping whole transformers apart gives 72.0%. Letting the model refuse low-confidence cases lifts accuracy inside the 589 records from 83.8% to 96.0% at half coverage, but on outside records it moves only from 55.5% to about 62%. A few records from the new utility help much more than any such trick: 25 local records alone give 65%, more than the 589-record model plus 25 local records (54%). Conformal prediction sets promised 90% coverage and delivered 62% on the new utility. All numbers were produced with MATLAB scripts that are in the repository. I make no claim of deployment readiness. The contribution is a measured account of how much of a reported DGA accuracy comes from the test design.

Keywords: predictive maintenance, power transformer, dissolved gas analysis, fault diagnosis, random forest, Duval Triangle, evaluation, domain shift.

1. Introduction

1.1 Why transformers need predictive maintenance

Power transformers are among the most expensive and most important parts of an electrical network. A transformer does not usually fail without warning. Over time, heat, electrical stress and ageing of the insulation can start a fault, for example overheating or electrical discharge in the windings, the core or the

insulation. If the fault goes unnoticed, the transformer can fail, which means an outage for customers, a costly repair or replacement, and a safety risk for the staff nearby.

Predictive (condition-based) maintenance tries to find such a fault while it is still small, so that the repair can be planned. The usual way to look inside a running transformer is to test its oil. The classical rules for reading such a test are simple and fixed. They work in many cases but they are not always reliable: they use fixed limits, they give no answer for some records, and they do not learn from new data. Machine learning can learn the pattern from past records, and this is why it has become popular for the task. Whether it is more reliable in practice is the question of this paper. The paper deals with the diagnosis step of predictive maintenance, which is naming the type of fault from a gas test. It does not forecast how long a transformer will last.

1.2 Dissolved gas analysis for readers outside the field

A power transformer is filled with insulating oil. When something goes wrong inside, for example a spark or a hot spot, the oil and the paper break down and release gases that dissolve in the oil. A technician takes an oil sample and a laboratory measures how much of each gas is present. This is dissolved gas analysis (DGA). This study uses five gases: hydrogen (H2), methane (CH4), ethane (C2H6), ethylene (C2H4) and acetylene (C2H2). Different faults make different mixtures. Acetylene points to arcing, ethylene to high temperature, and a lot of hydrogen with little else to partial discharge.

The standard IEC 60599 names six fault types: partial discharge (PD), low-energy discharge (D1), high-energy discharge (D2), and thermal faults below 300 C (T1), from 300 to 700 C (T2) and above 700 C (T3). The task for the model is to read five numbers and name one of these six.

Two older, rule-based methods are used as yardsticks. The Duval Triangle turns the shares of CH4, C2H4 and C2H2 into a point on a triangle that is divided into zones. The IEC ratio table compares three gas ratios with fixed limits. Both give no answer for some records. In this paper a record with no answer counts as wrong for the Duval Triangle, and the IEC table is reported both ways.

1.3 Problem and research questions

Predictive maintenance is only as good as the diagnosis behind it. Reports on machine learning for DGA often give accuracies of 90% or more. Such a number depends on where the records came from, how the labels were made, and how the data were split. A reader cannot tell this from the number alone. This paper asks four questions on public data:

- Which of four common classifiers is best on the standard 589-record set, and how does it compare with the Duval Triangle and the IEC ratio table?
- Why is the best accuracy near 83% and not near 95%, and which classes does the model confuse?
- How much does the test design (label origin, leakage, single source, same transformer in training and test) change the same model's score?
- What helps when the data come from a new utility: a confidence threshold, a few local records, or prediction sets?

1.4 Contribution

The contribution is a measured account, with code, of how far a reported DGA accuracy depends on the test design, and what a small amount of local data can recover. The model comparison itself is conventional.

2. Related work

Classical interpretation rests on the IEC 60599 standard and on the IEC TC 10 database of cases confirmed by inspection, described by Duval and dePablo (2001). The Duval Triangle comes from that work. The DGALab package of Ibrahim, Ghoneim and Taha (2018) implements the classical methods and ships labelled data sets. Nanfak et al. (2021) propose a ratio-based method built from labelled fault data and test it on 117 samples and on the TC 10 database, reporting fault-type accuracy of 97.4% for their method and 98.3% for the Duval Triangle. Machine learning studies on DGA often report single-data-set accuracies in the same range. This paper does not survey that literature, and it does not review predictive maintenance beyond the diagnosis step. It shows, with one model and public data, which test designs produce such numbers (Section 5.3), and it uses the corrected paired t-test for repeated cross-validation of Nadeau and Bengio (2003) and the SMOTE method of Chawla et al. (2002).

3. Data

The main data set has 589 records with the five gases and a fault label (data/DGA_transformer_dataset.csv). The class counts are D2 149, T1 111, T3 104, D1 91, PD 74 and T2 60. A provenance check against the public DGALab package (Ibrahim, Ghoneim and Taha, 2018) shows that these rows come from the DGALab file dataset_(589).xlsx. They are a mix of sources, and 115 of them match cases from the IEC TC 10 database, whose labels come from inspecting the transformer after the fault (Duval and dePablo, 2001).

Labels come from different practices, and this matters for the results. TC 10 labels follow physical inspection. Other labels were assigned by experts or by a rule. A set of 465 further records (data/picl_records.csv, from a public file of labelled records) is used as an outside test. It keeps only records that are confirmed, not duplicates of the 589, and above the IEC attention level, so that every record shows a real gas build-up. The 465 records come from 189 transformers or cases, 351 of them from one utility (NCEPR, 107 transformers).

Fujian records in the public file were labelled by a rule and are not used as ground truth.

4. Methods

Features. Each record gives 15 numbers: the logarithm of each gas (log of 1 plus ppm), six gas ratios, the total gas, and the three Duval percentages. Ratios use a small constant to avoid division by zero.

Models. A random forest (300 trees, 5 features tried per split, leaf size 1), a k-nearest-neighbour classifier, a support vector machine with an RBF kernel, and a two-layer neural network. Class imbalance is handled with SMOTE, written in MATLAB and applied only to the training part of each split (Chawla et al., 2002).

Splits. Section 5.1 uses a single 80/20 split with 117 test records. Sections 5.2 to 5.8 use repeated stratified cross-validation, or transformer-wise splits where stated, and report the mean and spread over folds. The test-design comparisons use one forest with fixed settings, so that differences come from the test and not from tuning.

Statistics. Where two models are compared over cross-validation folds, DGA_Rigorous_Eval.m contains the corrected paired t-test of Nadeau and Bengio (2003) and Wilson intervals for proportions. With 116 or 117 test cases, intervals are wide, and a gap of a few cases is not a real difference.

5. Results

5.1 The four-model comparison

Model	Accuracy	Precision	Recall	F1
Random forest	83.76%	84.12%	83.32%	83.44%
KNN (K = 1)	76.07%	76.17%	75.28%	75.53%
SVM (RBF)	76.07%	75.58%	74.19%	74.38%
Neural network	70.94%	71.71%	69.70%	69.70%

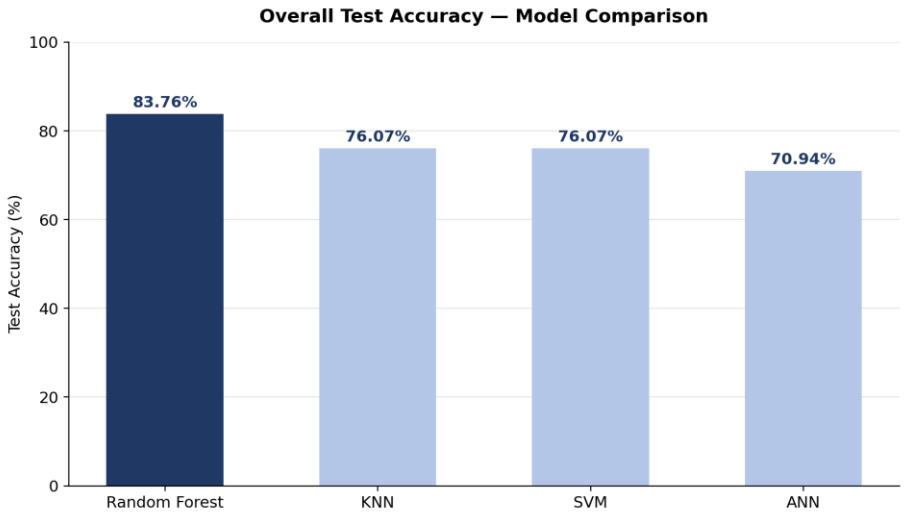

Figure 1. Test accuracy of the four models on one 80/20 split (117 test records).

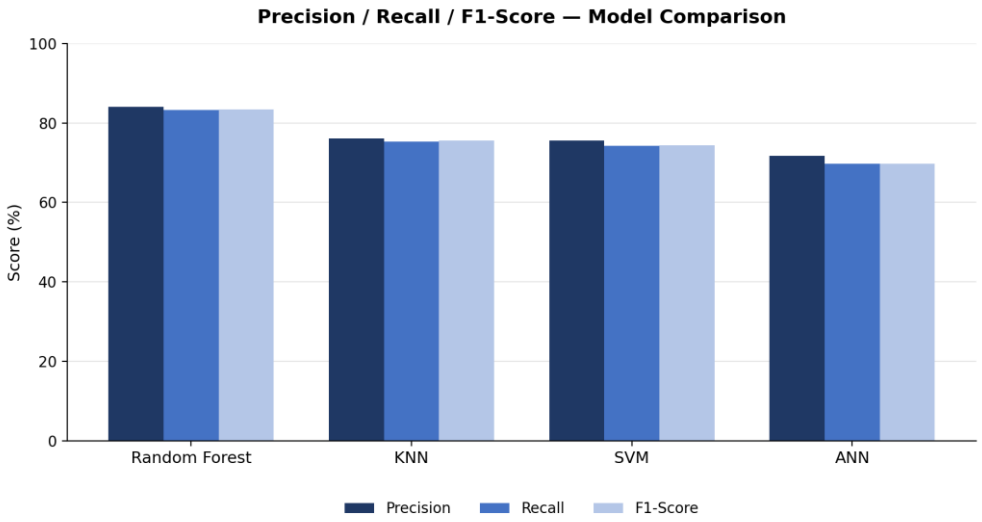

Figure 2. Macro precision, recall and F1-score of the four models on the same split.

Rule-based yardstick (no training)	Accuracy on all 589 rows	Comment
Duval Triangle (my simplified polygons)	62.8%	A record with no zone counts as wrong.
IEC ratio table	46.9%	Answers 64.3% of records; 72.8% correct on those it answers.

These are from one 80/20 split (117 test records) and come from DGA_Model_Comparison.m in code/. On 117 records one extra correct case is 0.85 percentage points, so the ordering between KNN and SVM is not meaningful, and the forest is ahead of the neural network by about 15 cases. KNN selected K = 1 by cross-validation, which means it memorises its training data, so its modest score should not be read as a limit of the method. The confusion matrix of each model is in figures/ (rf_confmat.png, knn_confmat.png, svm_confmat.png, ann_confmat.png). The comparison uses one split, so it is the weakest evidence in the paper. A repeated nested comparison of the four models is built into DGA_Rigorous_Eval.m (RUN_NESTED = true) and has not been run for this study. The sections below do not depend on this ordering. The rule-based yardsticks need no training, so their numbers are on all 589 rows and are not directly comparable with the 117-record split.

5.2 Why the accuracy is 83% and not 95%

A reader who sees papers with 95% to 99% on similar data will ask why this one is lower. There are four reasons, and each can be checked.

- The classes are cut from a continuous scale. Gas mixtures change smoothly with temperature and with discharge energy. The border between T2 and T3 is 700 C, and between D1 and D2 it is the discharge energy. Records near a border can fall on either side, and no model can separate them from the five gas values alone. In the cross-validated results, D1 (69.2%) and PD (72.5%) have the lowest recall, followed by T2 (78.3%). T3, the hottest and clearest class, reaches 90.7%.
- The labels are not all made the same way. Inspection-confirmed cases, expert opinions and rule outputs are mixed. A model has to learn all three habits at once.
- The data are small. 589 records over six classes leave 60 to 150 per class, and the folds have a spread of about 2 to 4 points.
- Most higher numbers can be reproduced by changing the test, not the model (Section 5.3).

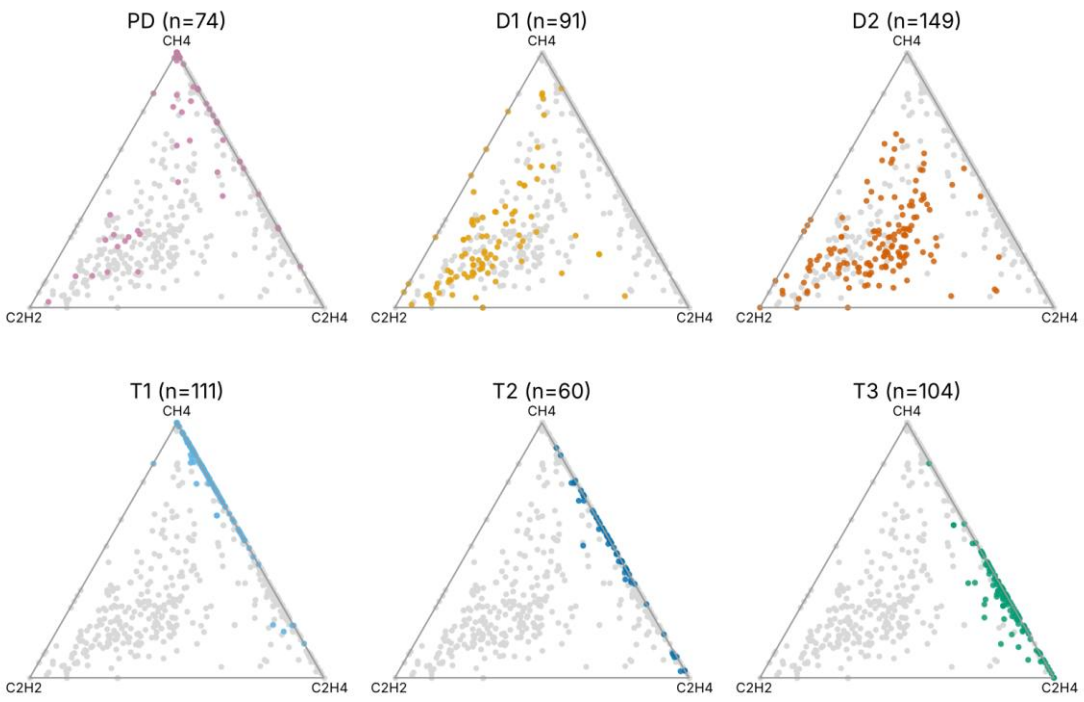

Figure 3. The 589 training records placed on the Duval Triangle coordinates (shares of CH4, C2H4 and C2H2), one panel per labelled class. Grey points are all records. The classes overlap, most of all D1 with D2, and the three thermal classes lie along one edge.

Figure 3 shows the overlap directly. D1 and D2 occupy almost the same region of the triangle. The three thermal classes all lie along the edge where acetylene is near zero and differ only in how far along the edge they sit, so T1, T2 and T3 are separated by a position, not by a gap. Partial discharge records are spread over the methane corner and the lower left, where they meet other classes. No classifier can split regions that the data themselves share.

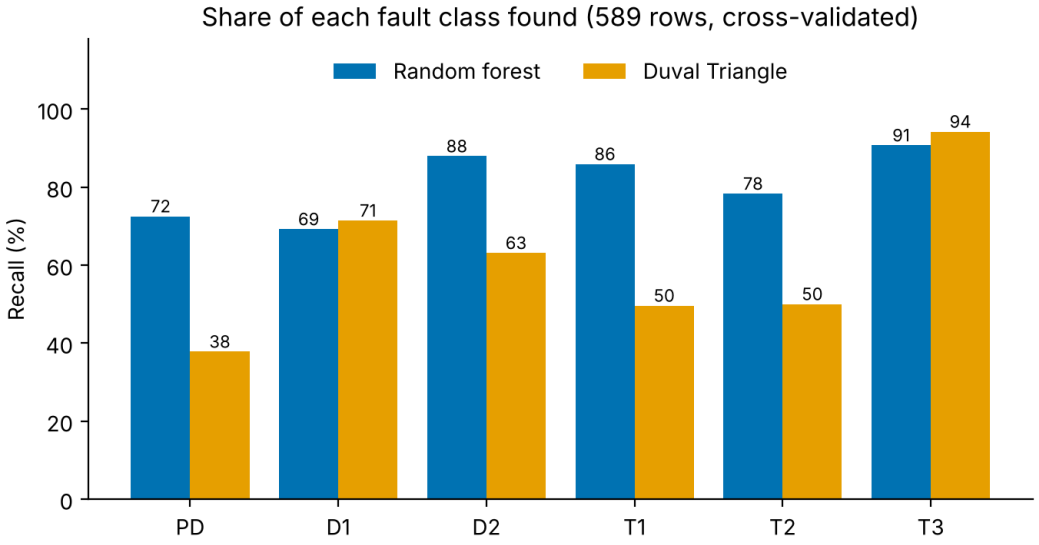

Figure 4. Share of each class found by the random forest and by the Duval Triangle on the 589 records. Duval scores 94.2% on T3 and 71.4% on D1, slightly above the forest for D1.

Figure 4 also shows where the forest adds value over the Duval Triangle: PD (72.5% against 37.8%), T1 (85.9% against 49.6%) and T2 (78.3% against 50.0%). Overall the forest scores 82.2% in this pooled count against 62.8% for the Duval Triangle on all rows. The Duval Triangle is better only for D1 and T3.

On the 116 TC 10 cases the order flips. The forest trained on the remaining 474 rows scores 74.5% on six classes (91.9% when only the three types PD, discharge and thermal are counted), and the Duval Triangle scores 87.9% (97.4%). This does not show that a rule beats learning. The Duval zones were drawn from cases like these, so the test favours the rule. The IEC ratio table scores 51.7% (80.2% by type).

5.3 How the test design moves the number

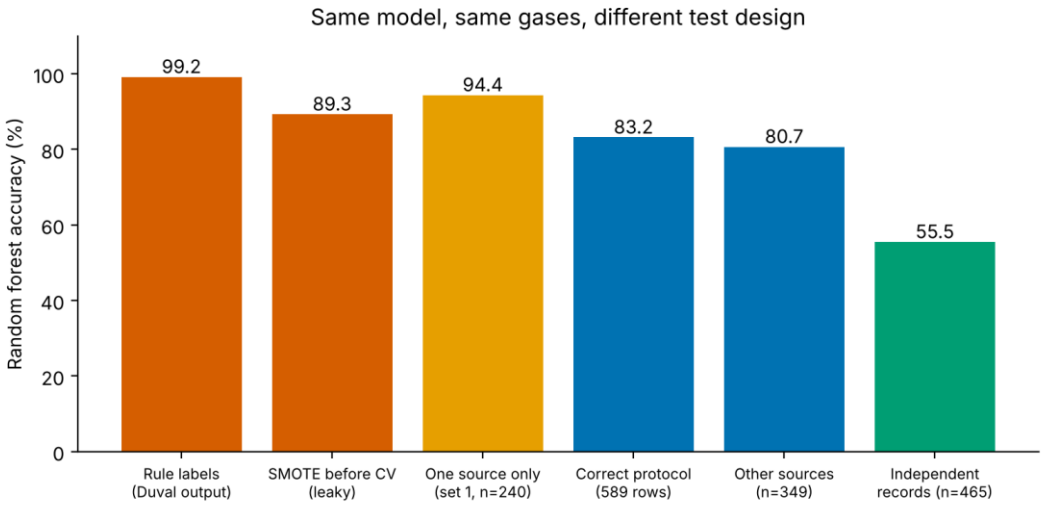

Figure 5. One random forest, five test designs. Orange and red bars use designs that overstate performance.

Test design	n	Accuracy	What is wrong or right with it
Labels taken from the Duval Triangle output	563	99.2%	The model learns a rule it is then graded against. Circular.
SMOTE applied before cross-validation	894	89.3%	Synthetic copies of test records sit in the training set. Leaky.
One source only (DGALab set 1)	240	94.4%	One laboratory, one habit of labelling. Easy.
Clean protocol, pooled rows	589	83.2%	Reference.
All other sources, without set 1	349	80.7%	Same protocol on the rest.
465 independent records	465	55.5%	Different utilities, no overlap with training.

The first row also explains why a thesis can report 99%: if the label is the output of a diagnostic rule, a flexible model will copy the rule. The point is not that high numbers are wrong. A number has no meaning without the label origin and the split.

5.4 The same transformer appears in training and test

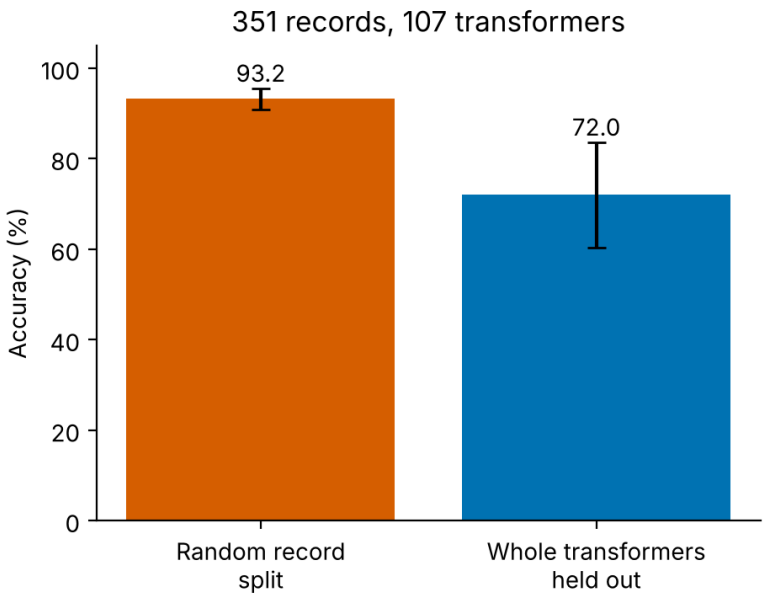

Figure 6. Inside the NCEPR utility, cross-validation by record against cross-validation by transformer.

NCEPR has 351 records from 107 transformers, so one transformer gives several records over time. A random split puts records of the same transformer on both sides and gives 93.2% (spread 2.3 points). Keeping each transformer entirely in either training or test gives 72.0% (spread 11.6 points). The model has partly learned the transformers, not the faults. The much larger spread also says that with few transformers, the result depends on which ones are in the test set.

5.5 Outside records: where the 83% goes

Trained on the 589 rows and tested on the 465 independent records, the forest scores 55.5%. The Duval Triangle on the NCEPR test sets of Section 5.7 scores 47.6%, so the forest is ahead but far from the 83% seen inside its own data. Two things are different outside: the gas levels and mixtures of other utilities, and their labelling habits. The independent set is also harder in itself, because it contains many D2 and T3 records and few PD and T1 records. A separate Python check, used only to test the MATLAB code, found the main mistakes to be T3 read as T2, D1 read as PD, and D2 read as D1. Those are the neighbouring-class errors expected from Section 5.2.

5.6 Can the model say when it does not know?

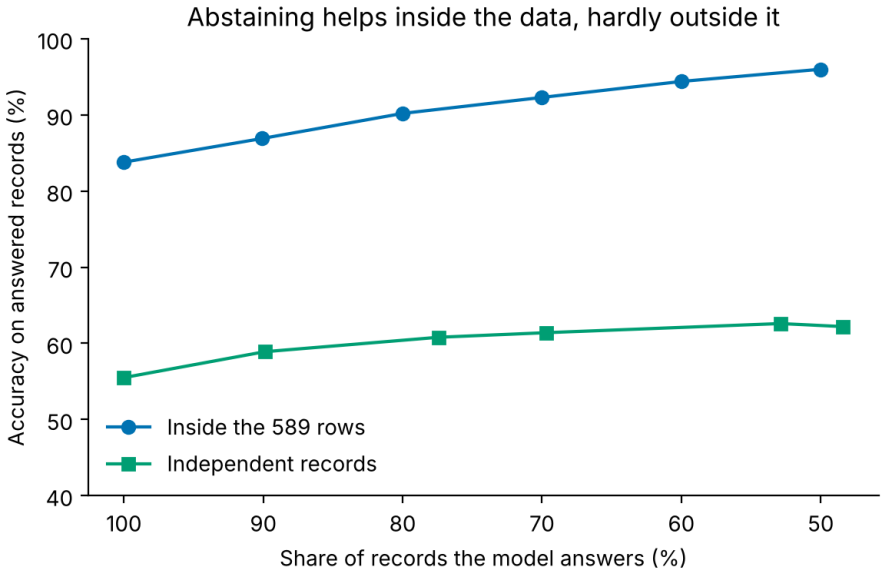

Figure 7. Accuracy on the records the model agrees to answer, for several coverage levels.

The forest gives a confidence from its vote share. If it answers only the 50% most confident records inside the 589 rows, accuracy rises from 83.8% to 96.0%. On the independent records, answering the 48% most confident ones gives 62.2%. The confidence learned inside the data does not carry over, because the model is confident about the wrong class when the data come from elsewhere. A confidence threshold set on one utility should not be trusted on another.

5.7 What a few local records do

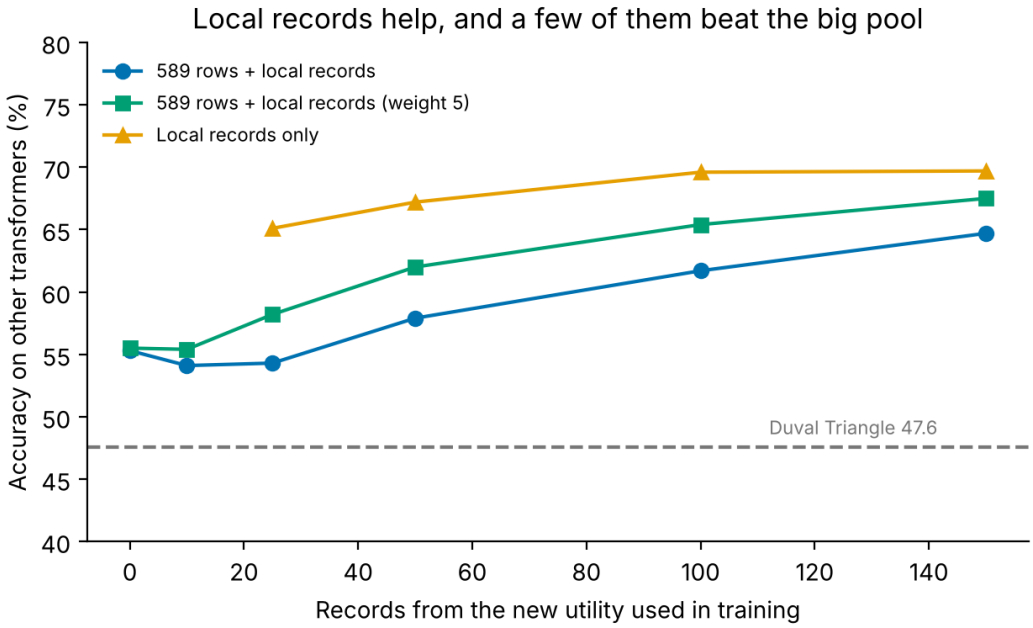

Figure 8. Accuracy on held-out transformers of NCEPR as local records are added. Mean of 20 repetitions; spread is 8 to 11 points.

Local records	589 rows plus local	589 rows plus local, weight 5	Local only
0	55.3%	55.5%	n/a
25	54.3%	58.2%	65.1%
50	57.9%	62.0%	67.2%
100	61.7%	65.4%	69.6%
150	64.7%	67.5%	69.7%

The test transformers are never in the local training records. Adding 25 local records to the 589 does nothing (54.3% against 55.3%), because 25 rows are lost among 589. Weighting them five times, or using them alone, helps clearly. 25 local records alone beat the 589-row model by about 10 points. The result is for one utility and the spread over repetitions (8 to 11 points) is large, so the order of the three curves is more reliable than any single value.

5.8 Prediction sets

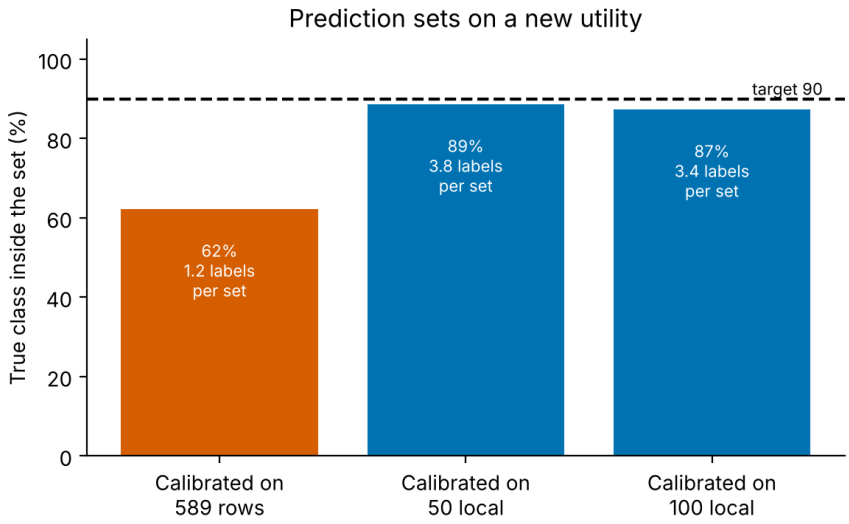

Figure 9. Conformal prediction sets aimed at 90% coverage.

Conformal prediction gives a set of labels instead of one label and promises the true label is inside the set 90% of the time, if calibration and test data are alike. Calibrated on the 589 rows, the sets contain the truth for only 62.2% of NCEPR records (average set size 1.24). Calibrated on 50 or 100 local records, coverage returns to 88.6% and 87.3%, but the sets then hold 3.8 and 3.4 of the six labels on average. That is a correct but weak statement. Both facts are consistent with the data not being exchangeable between utilities.

6. Discussion

The forest was the best of the four models on one split, which agrees with the idea that tree models suit this task: the Duval and IEC rules are themselves sets of thresholds on gas ratios, and trees learn such

thresholds without feature scaling. I did not test this explanation. What the audit adds is a sense of scale. The gap between the forest and the other models (about 8 to 13 points on one split) is smaller than the gap between a clean test and a test with rule-made labels (about 16 points), and far smaller than the gap between the 589 rows and outside records (about 28 points).

Three practical points follow. First, a DGA accuracy should be reported with the origin of the labels, the number of sources and whether records of one transformer can fall on both sides of the split. Second, a confidence threshold tuned on one utility does not carry over to another, and prediction sets that look sound on the training source can fail on a new one. Third, the most useful single step on a new utility was to collect a few dozen labelled local records, at least in the NCEPR experiment, which is one utility and therefore weak evidence for the general case.

7. Limitations

- The four-model comparison in Section 5.1 is a single split. A repeated nested comparison is in DGA_Rigorous_Eval.m (switch RUN_NESTED) but was not run for this study.
- The independent set is partly a public compilation. Its labels were not re-checked by me, and 351 of 465 records come from one utility.
- The Duval polygons are my own simplified version. Edge cases may differ from published implementations.
- Everything is a six-class problem on five gases. CO, CO2, gas rates over time and transformer type are not used, and real practice uses them.
- Nothing here was tested on a transformer in service. The numbers describe public records only.

8. Conclusion

This paper claims that, on public data, a random forest on five gases is about 82 to 84% accurate under a clean protocol and about 55% on records from other sources, and that high figures are easy to produce through label origin, leakage and single sources. It also claims that, in the experiments here, a small amount of local data helped more than a confidence threshold set on another source. It does not claim that the forest is better than the Duval Triangle in practice (on the inspection-confirmed TC 10 cases the Duval Triangle scored higher), that it works in service, or that it is the first study to see these effects. Similar warnings about leakage and shift exist in the general machine learning literature.

For predictive maintenance, the message is practical. A machine-learning reading of a gas test is a useful second opinion next to the classical rules, and on mixed data it found more faults than the Duval Triangle. It should not be trusted blindly at a new utility until it has been checked on that utility's own labelled records. In these experiments, a few dozen local records were what improved the model on a new utility.

9. Reproducing the numbers and availability of code and data

Open MATLAB with the Statistics and Machine Learning Toolbox and the Deep Learning Toolbox, put the files of data/ next to the scripts in code/, and run DGA_Rigorous_Eval.m and then DGA_Extended_Experiments.m. The second one prints the tables of Sections 5.2 and 5.4 to 5.8 and writes them as CSV files. Run times are about 30 minutes. Random seeds are fixed, but a different MATLAB version can change a result by about one point, as the two runs in

results/matlab_rigorous_run_output.txt and results/matlab_extended_run_output.txt show (independent accuracy 56.1% in one run and 55.5% in the other). The figures were drawn from the printed MATLAB numbers with a plotting program.

Code, data and results are in the repository github.com/kwadwoadugyamfi82-web/dga-transformer-fault-diagnosis-rf. The data files keep the terms of their original sources (data/README.md).

Declarations

Data and code availability. Code, data and results: github.com/kwadwoadugyamfi82-web/dga-transformer-fault-diagnosis-rf. The data files keep the terms of their original sources.

Competing interests. The author declares no competing interests.

Funding. This work received no external funding.

Acknowledgements. The project was supervised by Dr. E. K. Anto at KNUST.

References

IEC 60599: Mineral oil-filled electrical equipment in service. Guidance on the interpretation of dissolved and free gases analysis. International Electrotechnical Commission.

M. Duval and A. dePablo, "Interpretation of gas-in-oil analysis using new IEC publication 60599 and IEC TC 10 databases," IEEE Electrical Insulation Magazine, vol. 17, no. 2, pp. 31-41, 2001.

A. Nanfak, S. Eke, C. H. Kom, R. Mouangue and I. Fofana, "Interpreting dissolved gases in transformer oil: A new method based on the analysis of labelled fault data," IET Generation, Transmission & Distribution, 2021, doi: 10.1049/gtd2.12239.

A. Ibrahim, M. Ghoneim and I. Taha, "DGALab: an extensible software implementation for DGA," IET Generation, Transmission & Distribution, 2018, doi: 10.1049/iet-gtd.2018.5564.

L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5-32, 2001.

N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

C. Nadeau and Y. Bengio, "Inference for the generalization error," Machine Learning, vol. 52, pp. 239-281, 2003.

A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," arXiv:2107.07511, 2021.