Optimal Gas Subset Selection for Dissolved Gas Analysis in Power Transformers

José Pinto[1], Vitor Esteves[2], Sérgio Tavares[3], and Ricardo Sousa[4]

[1,2,4] *LIAAD-INESC TEC, Porto, Porto, 4200-465, Country*
jose.f.pinto@inesctec.pt
vitor.m.esteves@inesctec.pt
ricardo.t.sousa@inesctec.pt

[3] *Efacec SA, Moreira, Maia, 4470-605, Portugal*
smotavares@ua.pt

Abstract

The power transformer is one of the key components of any electrical grid and, as such, modern day industrialization activities require constant usage of the asset. This increases the possibility of failures and can potentially diminish the lifespan of a power transformer. Dissolved Gas Analysis (DGA) is a technique developed to quantify the existence of hydrocarbon gases in the content of the power transformer oil which in turn can indicate the presence of faults. Since this process requires different chemical analyses for each type of gas, the overall cost of the operation increases with the number of gases. Thus, a machine learning methodology was defined to meet two simultaneous objectives, identify gas subsets and predict the remaining gases, thus restoring them. Two subsets of equal or smaller size to those used by traditional methods (Duval's triangle, Roger's ratio, IEC table) were identified, while showing potentially superior performance. The models restored the discarded gases, and the restored set was compared with the original set in a variety of validation tasks.

1. Introduction

The energy sector is one the most competitive markets, having a global reach (Georgilakis, Katsigiannis, Valavanis, & Souflaris, 2006), where the strife for innovation and efficiency improvements is ever present. A global trend toward deregulation and privatization (Gavrilovs & Vītoliņa, 2011; Mao, 2000), has continuously intensified the existing competition in the last few decades.

From all components forming the power systems landscape, the Power Transformer (PT) is one of the most important and common (Georgilakis et al., 2006). Due to this increase in competition, decisions to increase PT loads have become a common occurrence (Gavrilovs & Vītoliņa, 2011), which coupled with the reduction of maintenance activities and the fact that most PTs installed are reaching the end of their useful life (Zhao et al., 2017) greatly increased PT failure rates. With all these compounding factors, PT failures, even though they have extreme reliability, have experienced a worldwide increase (Zhao et al., 2017; Georgilakis et al., 2006).

These failures have a multitude of consequences. From the expected, which include transmission and distribution interruptions, or even wide area blackouts (Dhonge, Swamin, & Thosar, 2015; Nurmanova et al., 2020). To the more unexpected, such as damage to the PT and grid, environmental damage, fire or in extreme cases explosions (Georgilakis et al., 2006; Nurmanova et al., 2020). In case of extreme failures, damage to only the PT might be the best case, but due to its high cost (accounting for as much as 60% in power stations (Sarajcev, Jakus, & Vasilj, 2020)), this is still very disruptive.

Like the consequences, the causes of PT failure are diverse. They can be external to the PT, such as weather events (wind, lightning, and snow); human error, including poor design, transport, or installation; or extreme events such as fires or earthquakes (Gavrilovs & Vītoliņa, 2011; Nurmanova et al., 2020). However, in practice, most failures are internal (as much as 80% (Chauhan & Sinha, 2015)). They include short circuits, overloading, insulation failure, and others (Georgilakis et al., 2006).

With how damaging PT failures can be, it is no surprise that a variety of mechanisms to warn the operator, impede failures or stop the PT operation are commonplace in currently operating devices. Coupled with this variety of safety mechanisms, maintenance is always performed, with traditional preventive maintenance and testing being due to costs (Ravi, Drus, & Krishnan, 2019; Gavrilovs & Vītoliņa, 2011), fazed out as more intelligent approaches become more prevalent.

Given the impact of transformer faults, it becomes clear why the most frequently tackled problem related to PT monitoring and management is fault diagnosis or prediction, where the objective is to find the causes behind an already occurred fault.

For this problem, various methods exist, but by far the most common involves the use of Dissolved Gas Analysis (DGA) data (Ravi et al., 2019; Dhonge et al., 2015), where the concentration in oil of various gases is measured, which are then used for diagnosis in a variety of techniques (Naresh, Sharma, & Vashisth, 2008; Dhonge et al., 2015; Mirowski & LeCun, 2012).

Remaining Useful Life (RUL) prediction and Health Index (HI) obtention are two closely related problems that also show some prevalence in this domain. The prediction of RUL entails determining how much time remains until a PT will become nonoperational, while HI obtention aims at obtaining a single number or rating that accurately describes the PT health in terms of how long it will likely last (Velásquez, Lara, & Melgar, 2019; Sarajcev et al., 2020). The results obtained from these problems can then be utilized to define maintenance strategies or to better prepare for an eventual failure (Velásquez et al., 2019).

Despite the importance of PTs to the grid, there are still many problems that show almost no prevalence in literature. One, proposed by Efacec, was found to be of great importance, which entails identifying small, yet viable, DGA subsets that can be effectively employed, without affecting performance, by a variety of classical and Machine Learning (ML) based methods.

A solution to this problem would not only decrease costs and time expenditure, but also improve maintenance actions. Each gas measured incurs a cost in terms of expendable materials, personnel, expensive monitoring equipment, and increased down time, while augmenting the margin for human error, thus possibly reducing PT lifespan.

To effectively tackle this problem, Efacec provided a proprietary and specific dataset, which will be the focus for most of the presented work. Despite great efforts to identify other datasets to complement this one, no publicly available contained the immaculate DGA variables required for this work and no private dataset could be reached.

To accomplish this objective, two goals were defined. The first is to empirically determine the most important gases, something that to the authors' knowledge has never been attempted (non-empirical studies based on transformer operations, chemistry, and other factors are known.). The second goal is to develop ML based methodologies, which can be used to predict the values of the remaining gases (were these measured) and thus complete the gas set.

This methodology does not aim to solve a particular problem, such as failure diagnosis. Instead, it first aims to guide the PT monitoring process, aiding in the selection of sensors and measured PT attributes. Second, it is intended to act as a middleman between the reduced gas samples and the models, such as fault diagnosis models, which may require more or different sets of gases.

The main idea behind our approach is that gas relations in PT oil are complex; thus, if a model could capture these, then that model could be employed to predict the values that were otherwise not recorded, thus letting DGA-based methods operate with only a fraction of the required gases actually needing to be measured.

Although there is a possibility that some advanced ML methods could implicitly identify gas relations, even if some are not present, it is unlikely. Furthermore, the same is not possible for classical and other already developed methods that require specific sets of gases, which may not be available.

Therefore, to achieve our ambitious goals, a complete data science pipeline, including visualization, preprocessing, modeling, and fine-tuning operations is presented, with the intent of making the work as comprehensive, understandable and replicable as possible, brought on by the proprietary nature of the dataset and code base.

The remainder of this paper is structured as follows. In section 2, related work in PT failure diagnosis is presented. Then, in section 3, a thorough explanation of the dataset and visualization approach is done. The detailing of the modeling methodology and validation tasks is presented in section 4. In section 5, an overview of the final results for the regression and validation tasks is performed. Finally, the main conclusions, contributions, limitations, and future work are presented in section 6.

2. RELATED WORK

In our work, there are some areas of particular importance. We will start by looking at works pertaining to DGA based fault diagnosis techniques, with a special focus on classical approaches. We will follow this with works relating to subset feature selection. Finally, missing value imputation works will be analyzed.

At the core of our subject problem and, of course, our methodology, DGA takes center stage. Thus, a good understanding of its role in the overall ecosystem is crucial, not only for the guidance of the developed work, but also for the reader to properly understand our contributions. Of special importance are the classical methods (Duval's triangle (Duval, 1989), Roger's ratio (Ward, 2003), IEC table (Duval, 1989) and key gas method (Londo, Çelo, & Bualoti, 2015)), which are critical for our results validation methodology.

Dhonge et al. (Dhonge et al., 2015) developed a summary of different classical deterministic DGA failure diagnosis techniques, such as Duval's triangle, Roger's ratio and IEC ratios. A comparison between Duval's triangle and a state of the art ANN is performed, with the results showing the superiority of the ANN for the task.

Adrianto et al. (Andrianto et al., 2020), in their work, employed Markov models to investigate PT oil reliability and availability, performing a comparison to widely used methods for this task. Using a dataset containing DGA and breakdown voltages, the new method achieved 85% accuracy, improving on the second best model, Duval's triangle, which achieved only 61%. Thus, the new method was found to be a significant improvement over current alternatives and Total Dissolved Combustible Gas (TDCG) to be of particular importance.

With the objective of improving the International Electrotechnical Commission (IEC) table for DGA fault diagnosis, while providing understandable results derived from data-driven models, Miranda and Castro (Miranda & Castro, 2005) present their work, where Transparent Fuzzy Rule Extraction from Neural Networks (TFRENN) was developed; a combination of Artificial Neural Networks (ANN) and a Fuzzy Inference System (FIS). Two variants of the IEC table, as well as some combinations of ANNs and (FIS) were tested, from which the new method attained the best results, while being more broadly applicable and providing confidence values.

The fault diagnosis of PTs is a mature field, and thus there are, of course, many works that focus solely on furthering the ML field within this domain, with little mention of the classical methods.

Mirowski and LeCun (Mirowski & LeCun, 2012) conducted a review of DGA ML techniques, consisting of 15 different methods applied to two distinct problems, in 2012. One problem entails the binary decision of whether a transformer will or not fail in the near future. The other is related, turning the binary problem into the regression of time until failure. The dataset used consists of DGA data and PT characteristics, such as voltage, power, and age. From the large number of methods, including Decision Trees (DT), Support Vector Machines (SVM), Local Linear Regression (LLR), ANNs and K Nearest Neighbors (KNN); SVMs were found to be the best for the first task, while LLR for the second, with ANNs coming in second for both instances.

Naresh et al. (Naresh et al., 2008) proposed a new neural fuzzy network approach for DGA data interpretation. A combination of competitive learning and subtractive clustering was utilized for variable selection and fuzzy rule base creation, before being fed into the network. The new approach was compared to a variety of methods, including Roger's ratio, radial basis function neural network, and Fuzzy C-Means (FCM) in two different datasets. The new proposed approach generally achieved the best results.

For the development of our proposed methodology there are other fields of particular importance. One is feature selection, which within our work we preferred to refer to as subset selection, due to differences in the overall goals and application. Some of the gathered works focus on applications in the PT domain, while others focus on the task itself (feature selection).

An interpolation method for prediction of short-circuit severity was developed by Nurmanova et al. (Nurmanova et al., 2020), with piecewise Hermite interpolation, linear interpolation, and natural cubic spline being tested on Frequency Response Analysis (FRA) data. Feature selection was done using Sequential Forward Selection (SFS) - also known as Greedy Forward Search (GFS), amongst other names - with cubic Hermite interpolation attaining the best results.

Sun et al. developed an extensive enumeration and review of data-driven fault diagnosis methods (Sun, Huang, & Huang, 2012). Models including SVM, Particle Swarm Optimization (PSO), Fuzzy Logic (FL) and Wavelet Network (WN) are presented, and their weaknesses, strengths, and use cases are detailed.

A new genetic algorithm-based methodology for feature subset selection is presented by Tan et al. (Tan, Fu, Zhang, & Bourgeois, 2008). This method is applied to the specific task of microarray gene expression data analysis with the goal of identifying the most important genes. This method employs multiple common feature extraction strategies to create the original population, while obtaining equivalent or superior results to any of these, showing greater generalization capabilities in multiple tested datasets.

Another important field is missing value imputation, which, due to our development of a new imputation method aiming at improving results, gains far more weight. Therefore, we will now provide an overview of some works relating to this task.

The work by Silva et al. (Silva-Ramírez, Pino-Mejías, López-Coello, & Cubiles-de-la Vega, 2011) presents a new MultiLayer Perceptron (MLP) based missing value imputation technique. This method works by imputing all missing values in all features simultaneously and was contrasted with several common methods. These methods are mean/mode, regression and hot deck, with the novel method achieving the best results, with particular success for categorical missing values.

A survey of missing value imputation, including missing value types, imputation methods, and imputation quality validation is presented by Liew et al. (Liew, Law, & Yan, 2011). Throughout this work, a focus is placed on gene expression data, in particular regarding domain-specific knowledge assisted approaches. However, the techniques are sufficiently general to be relevant to our target domain.

From this analysis of the related work, it was found that very simple deterministic methods such as Duval's triangle and Rogers ratio, are still very prevalent for DGA based fault diagnosis. Although empiric ML methods, in particular ANNs and SVMs, achieve better results, a lack of comparisons due to the unavailability of good publicly available datasets is a major limiting factor. Regarding both feature selection and missing value imputation techniques, most works in the PT domain do not detail or even state which methods are tested and employed. However, since both fields are quite vast, it is very easy to find a variety of high-quality works and methodologies; although one has to look outside the PT domain.

3. MATERIALS AND METHODS

The first task of our methodology involves exploring the supplied data set. Given its proprietary nature, a detailed description of its contents, variable distributions, and relations will be given to make the replication of our work as easy as possible.

It is important to note, to make the replication of our results easier, that all code was developed in the Python language, with the main libraries involved being pandas, numpy, and sklearn. Thus, in any case where implementation details or parameters are not mentioned, the default for these libraries should be assumed.

3.1. Dataset Description

With 1000 entries and 39 variables, by ML standards, the dataset is small, while still being more complete than that available to many authors in the field. 200 different transformers are present, with on average 5 entries per transformer.

An analysis of variable statistical moments was performed, were it was concluded that the variables have very different scales and distributions, thus requiring special care.

To better understand the distributions, a set of univariate graphical plots was created, from which most numerical variables, in particular DGA gases, were found to be very left-skewed. Similarly, the categorical variables were found to be very unbalanced, with a single class dominating.

This analysis was followed by a bivariate one, where groups of highly correlated variables were identified, which could prove crucial for the regression tasks. Another important finding was that the variance of some variables is related to the values of another. Finally, nonlinear relations were also spotted.

To better understand these non-linear relations, sets of 2 and 3 variable regressions (polynomial, exponential, logarithmic and cubic splines) were performed. Instances were multiple underlying models could be spotted on 2 variables were properly differentiated with the third one.

With the knowledge obtained from the visualization approach, we proceeded with the preprocessing steps. The first of which is the handling of missing, incorrect, and conditional string values (those like ”<0.2” or ”>=8”). Almost all columns contain missing values, with their amount ranging from 0.88%s to 85.4%, with this leading to 98.1% of entries having at least one missing value.

The combination of a small dataset with many values that have to be corrected makes the simplest method of removing problematic rows unfeasible. As such, we will focus on the imputation process.

3.2. Imputation Techniques

Three imputation techniques we tested for numeric values, mean, regression and a novel method, Regression Sampling Imputation (RSI). A more in-depth explanation of the RSI method will be provided, for which a flowchart of the algorithm can be seen on image 1.

However, before, it bears mentioning that the conditional string imputation utilized the same methods, simply restricting the datasets beforehand to the conditional value. Categorical value imputation similarly employed mode, classification and Classification Sampling Imputation (CSI), which simply replaces the regression models with categorical ones.

For RSI and CSI, first the presence of missing values is verified, and if confirmed, a set of regression models is created utilising all subsets of a given small size, with the given variable as a target.

Models with performance under a selected threshold are discarded, with a fallback strategy of mean imputation in the case that no model remains. Then, for each missing value, a random model is sampled with its likelihood weighed by its performance, and the value is predicted.

Then whether the value lies inside the data range is verified, and in the negative case, the model is resampled up to a maximum of 10 times. If the limit is exceeded, mean imputation is instead used for this value.

There were a few goals behind the development of this method. The first is to preserve information better than those such as mean imputation. The second is to introduce a degree of variance and reduce bias in the predictions, unlike simple regression, which can have a lot of bias. And the third is to introduce this variance while preserving data distributions and relations better than simple random noise addition.

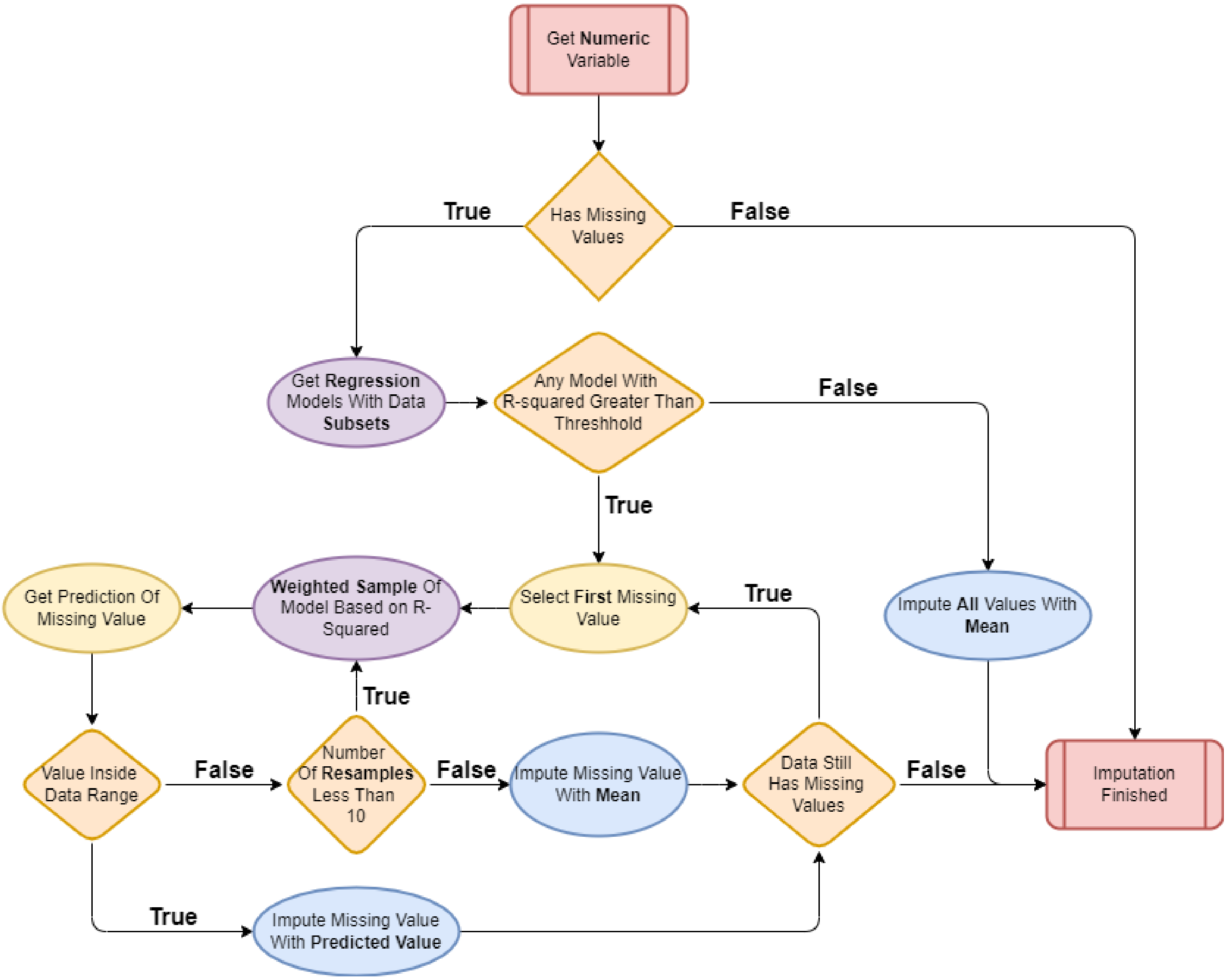

Get Numeric Variable
Has Missing Values
True
False
Get Regression Models With Data Subsets
Any Model With R-squared Greater Than Threshhold
False
True
Select First Missing Value
Weighted Sample Of Model Based on R-Squared
Get Prediction Of Missing Value
True
Impute All Values With Mean
Value Inside Data Range
False
Number Of Resamples Less Than 10
True
False
Impute Missing Value With Mean
Data Still Has Missing Values
False
Imputation Finished
True
Impute Missing Value With Predicted Value

Figure 1. Flowchart of the regression sampling imputation algorithm

3.3. Data transformations

With the missing values filled and the incorrect values corrected, a set of transformations followed. The first one, one hot encoding, was required by all models and does not have any tunable parameters, and as such, was always employed.

Due to the large number of variables, compounded by the one hot encoding process, dimensionality reduction methods were investigated, from which Principal Component Analysis (PCA) and extra trees importance were found to be promising.

The selection of variables to group for PCA was done by selecting a minimal correlation threshold after which the variables are combined. For the extra trees method, a regression is performed and the importance of each variable for the predictions is obtained. Variables with low importance are removed.

The problems that can occur due to very different distributions were addressed by Box-Cox transformations, approximating the variable distributions to normal. This transformation has the known side effect of improving training times for some methods.

To address the problem of different scales, normalization was employed, scaling the values to the range from 0 to 1 (variables originally assumed only positive values). Another possibility for this task was standardization, making the mean 0 and standard deviation 1. However, this second was not tested, as other preprocessing and modeling parameters were expected to have a greater impact on the results and thus took precedence.

It is important to note that since the DGA variables are the targets, they are not transformed during the preprocessing step. Any transformations to these variables only occur during the modeling stage in the cases were the variables are kept in the subset, and as such used as features.

3.4. Modeling

Now, we move on to modeling. The first component is the subset selection algorithm. Given the time complexity of the models we trained, only a very small set of all subsets could be tested and as such two simple subset selection methods were selected. The first is GFS, while the second is Greedy Backward Elimination (GBE). These methods work by adding or removing a variable at a time from the subset, in a greedy fashion. The addition/removal of each variable for each size is tested, with the best being selected.

Inside the subset selection loop, grid search cross-validation is used to select model hyperparameters, while a set aside set is used to choose between the models. For cross-validation, five folds were used, while the minimum of the R-squared for all predicted variables was used for comparison.

Four different model types were used, Support Vector Regressors (SVR), Multi Layer Perceptrons (MLP), Gradient Boosting Regressors (GBR) and Random Forests (RFs). Both GBRs and SVRs do not have innate multi-output capabilities. This was easily corrected by training a model for each output and using all as a group. The same was done while keeping the original, for MLPs and RFs, which do have multi-output capabilities, thus creating two variants for each. This led to the total of 6 models being tested.

4. METHODOLOGY CALIBRATION AND EVALUATION

Our approach contains a large number of parameters and settings that must be selected, including which preprocessing steps to apply, their parameters, modeling hyperparameter ranges, and others. Given this large amount and the computation time for each test, each parameter was selected independently.

4.1. Imputation Techniques Fine Tuning

Caused by computational constrains, this initial parameter tuning was performed via a visual analysis, comparing the imputed values to the original ones, while trying to minimize the discrepancy between the distributions in terms of peaks and statistical moments.

The first set of parameters tuned is that for value imputation. Both numerical and conditional string regression imputations were done by KNN. The number of neighbors and the distance metric are the parameters to be tuned, from which a k of 5 and Euclidean distance proved to be the best. Categorical classification imputation similarly utilized KNN; however, only a k of 1 was applicable and different distance metrics did not affect the results, thus remaining at the standard Euclidean distance.

RSI employed exponential, polynomial, logarithmic, and cubic spline models. All models for combinations of N variables are trained, as as such N has to be selected. A value of 3 was found to be optimal. Other methods could, of course, be added to the existing ones or replace them, but simpler ones are recommended, due to the large amount of models causing expensive computation, and to increase the variance similarly to usual ensemble methods.

Both exponential and logarithmic regression do not have any tunable hyperparameters, thus leaving polynomial and cubic spline. As it is unfeasible to tune the polynomial degree for each trained model, an automatic selection process needed to be selected. From a few tested, a F-test from an Analysis Of Variance (ANOVA) table was selected. While the maximum degree was tested and set to 10, for cubic spline, the number of knots, their locations, and the degree of each section have to be selected.

The minimum degree for the curve to be smooth, 3, was selected. The knots were placed at equidistant quantiles to ensure areas with more data points are better represented. Various numbers of knots were tested, with 3 achieving the best results.

4.2. Broad Hyperparameter Range Selection

With the imputation methods working, an initial preprocessing configuration was selected and the selection of coarse model hyperparameter ranges started. Each hyperparameter was added to the grid at a time in a greedy fashion, and the ranges were tuned so that the best hyperparameters found were not at any extremes. If a hyperparameter was found to have a negligible impact, it was discarded. With this, broad ranges of the most impactful hyperparameters were obtained.

It was surprising that some generally important parameters for MLPs were found to be insignificant, such as momentum, learning rate and regularization terms. Finally, it is important to mention that for both variants of MLPs and RFs, the selected hyperparameters were identical.

4.3. Transformation Fine Tuning

With reasonable hyperparameter ranges selected, we moved back to tuning the preprocessing steps. For this, the modeling results were used as a guideline, which requires a systematic way of comparing them. The minimum R-square can easily be used to compare results for subsets of the same size. However, one subset for each size (from 0 to 8) is created and a way to compare them was needed.

At this point, it bears reinforcing that two distinct problems were tackled, one using the whole dataset and another, for greater generality, using only the DGA data. The preprocessing steps were tuned individually for these.

As such, for the full dataset, the results were considered superior if the subset of minimum size achieving a R-squared above 0.7 was smaller, or in case the size was the same, if the value was greater. As the DGA only dataset obtained less dimensionality reduction and better results at high subset sizes, the same was done but with a threshold of 0.9.

With a consistent way to compare the results, we started by selecting the imputation methods from the set of 3 developed. In all cases, our methods, RSI or CSI, achieved better results. In this case, as there might have been a degree of bias towards the methods being of the same type, changing these as a group, rather than individually, was also tested. What this means is that changing all to regression and classification; or to mean and mode was tried. However, the same results remained.

Then, we moved to the PCA tuning, where the correlation threshold has to be selected. Totally disabling the transformation was also tested. For both datasets, a threshold of 0.8 was selected.

For extra trees dimensionality reduction, the decisions involve weather to disable it or, in case it is used, how many variables to remove. Disabling it achieved better results for both datasets.

The Box-Cox transformation only has the choice of whether to disable it. In this case, for the full dataset enabling it achieved better results, while for the DGA only dataset, the opposite was selected. Disabling Box-Cox had the downside of doubling training times.

Similarly, disabling normalization was tested, but with an increase in training time of more than 100 times, it was never possible to obtain the performance values and this option was discarded.

4.4. Hyperparameter Fine Tuning

The final step of our fine-tuning approach involves returning to the model hyperparameter ranges and changing them to a more detailed and constrained set to achieve the best possible results. This was done identically to the initial selection of the hyperparameter ranges.

4.5. Validation

At this point, we were ready to obtain the final results, with knowledge on how to interpret them, completed pre-processing and fully automated subset, model and hyperparameter search.

Still, the results obtained are in the form of regression results, which are not indicative of the actual performance in a task where these results would be applied. As such, a set of validation tasks was created.

Most of the methodologies that employ DGA data do so for fault diagnosis and detection, thus being linked to abnormal events and outliers.

As such, the first validation tasks focus on the analysis of outlier values. First, regression metrics are used to compare outliers and then binary classification metrics are used for the same goal. Thus, we first look at the actual outlier values and then at weather the same points are predicted as outliers. To find outliers in the first place, the outlier detection method from box plots is used. Then the set of points that are found in either the original or predicted data are obtained. Finally, these sets of points are compared.

After this, a second set of validation tasks were devised. These do not further compare the obtained regression results, but instead utilize common DGA failure diagnosis methodologies, with the predictions from the real data being compared to those obtained from the predicted data. These tasks should provide a better idea of the performance obtained in a real scenario. Given that each of these methods provides as output the prediction of the occurred fault, this analysis was approached as a multiclass classification problem.

A total of 4 methods were tested, Duval's triangle, IEC table, Rogers ratio, and key gas method. Unfortunately, the key gas method is less formal, relying on the expertise of the user, and as such, there seems to be no globally accepted standard for interpretation. Nevertheless, the values given by Londo and Çelo (Londo et al., 2015) were utilized for this analysis.

5. RESULTS

We will now thoroughly detail the obtained results for both datasets, not only in the original regression task, but also in the set of validation tasks that were outlined. We have divided this section into two, first presenting the results for the full dataset, and then those for the DGA only one.

Throughout this section, a large number of results tables are referenced, which were moved to the appendix section to avoid disrupting the structure.

5.1. Full Dataset

We remind the reader that the full dataset contains all the contents provided (except those that might have been removed by dimentionality reduction), such as sample and PT information, particle data and furanic compounds.

For this set of results, SVRs never achieved the best results, and Oxygen (O2) is never a part of any subset, thus being always predicted.

The general results' metrics can be seen on Table 1, with similar ones having been utilized for the selection of preprocessing steps and hyperparameter ranges. Here, a large amount of information is present, bearing some explanation. For each cell, a value of "N/A" indicates the presence of the variable in the subset. All other cells contain 3 separate performance metrics, R-squared, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Only R-squared was used to compare the regression results, but the others allow us to get a better sense of the error scale. Knowing this and by employing the R-squared threshold of 0.7, we verify that the number of gases can be reduced from 9 to 3 (Hydrogen (H2), Acetylene (C2H2) and Methane (CH4)).

Table 2 contains the metrics for regression on outliers, with the only difference to the previous set of results being the presence of ”nan” values, which appear if a variable does not have outliers. With an analysis of the ”worst” column, we verify that the results are very poor, whereupon further investigation, O2 was found as the culprit. The fact that all outlier values are very close together while being very far from the rest of the distribution, causes the R-squared to provide a very low value. However, when compared to the scale of the values, the error is quite small, and so, unless proper differentiation between outlier values is required, these results should not be a problem.

Table 3 shows the results for the binary classification on outliers, for which the structure is similar. Since this is now a classification problem, the metrics were replaced with accuracy, precision, recall, and F1. Here, the thresholds for accuracy of 0.9 and of 0.7 for all other metrics were selected. Given these, a minimum subset of size 3 (H2, C2H2 and CH4) can be achieved, which is much better than that for the regression on outlier metrics.

The results for the first problem validation task, using Duval’s triangle, are shown on Table 4. As this is now a multiclass problem, each column now represents a metric, with weighted and macro-averaged precision, recall, and F1, as well as accuracy being used. There are now 6 classes creating a harder problem, thus the thresholds were reduced to 0.6 and 0.4. Using these, the size of the subset can be reduced all the way to 1 (H2). It is important to note, regarding the perfect results for the subset of size 8, that these occur because the only variable being predicted, O2, is not used by Duval’s triangle, and thus, in both instances only real values are utilized.

The second validation task utilizes the IEC Table, for which the results, that are presented in an identical manner, can be seen on Table 5. With an increase to 9 classes, the thresholds are once again lowered to 0.55 and 0.35. Thus, a subset size of 2 (H2 and C2H2) is obtained. The perfect results for subset size 8 are caused by the same factors.

The validation task using Rogers ratio is the third one presented, with its results shown in Table 6. With a reduction of only 1 in the number of classes (from 9 to 8), the thresholds remain unchanged, for which a minimum subset size of 0 (only support variables) is achieved, with great results being attainable by using only support variables.

The final of our validation problems utilizes the key gas method. This is in theory the easiest of the problems, with only 4 classes, for which the results can be seen on table 7. With a decrease in the number of classes, the thresholds were increased to 0.7 and 0.45. Thus, a size of 1 (H2) is achievable. Despite this good performance, there is some concern regarding the instability of the results for larger subset sizes. This indicates that the values obtained for this method might be flawed. Despite this, as no information contradicting our implementation was found, these were kept.

5.2. DGA Only Dataset

The DGA only dataset utilizes only DGA gases to make predictions. In this section, the same set of validation approaches and selected thresholds are utilized. Thus, to reduce redundancy, only the most important differences or similarities will be addressed.

From Table 8 we can see that the performance for smaller subset sizes is lower than for the full dataset, while conversely, for higher sizes, it is equivalent or even superior. Thus, a minimum size of 6 is obtained.

The regression on outliers task shows very similar results, with O2, once again hindering the performance. While, the binary classification task achieves better performance, with a reduction to size 4 (C2H2, H2, C2H4, and N2). These are shown on Tables 9 and 10. One thing to note is the presence of ”nan” values outside of the Nitrogen (N2) column. These are caused by all the predictions being of the same class.

The results for the validation tasks involving Duval’s Triangle, IEC table, and Rogers ratio are very similar, with a minimum subset size of 3 (C2H2, H2 and C2H4) being obtained for all. These results for Duval’s Triangle are in Table 11. The other tables have not been included as the information is almost identical.

For the final validation task using the key gas method, found in Table 12, a subset size of 2 (C2H2 and H2) is obtained. This time, the results are far more stable than for the full dataset.

Throughout this section we presented a large amount of information in the form of metrics tables, which can be difficult and time consuming for the reader to fully understand. Therefore, in images 2 and 3 we present a summary of the most crucial information. From this, we can easily identify the poor results for Outlier Regression in both instances. We can also see that for most validation problems, both the full and the DGA only datasets, achieve the intended results for subsets of size 3. Finally, it is also clear that the results for key gas ratio are unstable in the full dataset.

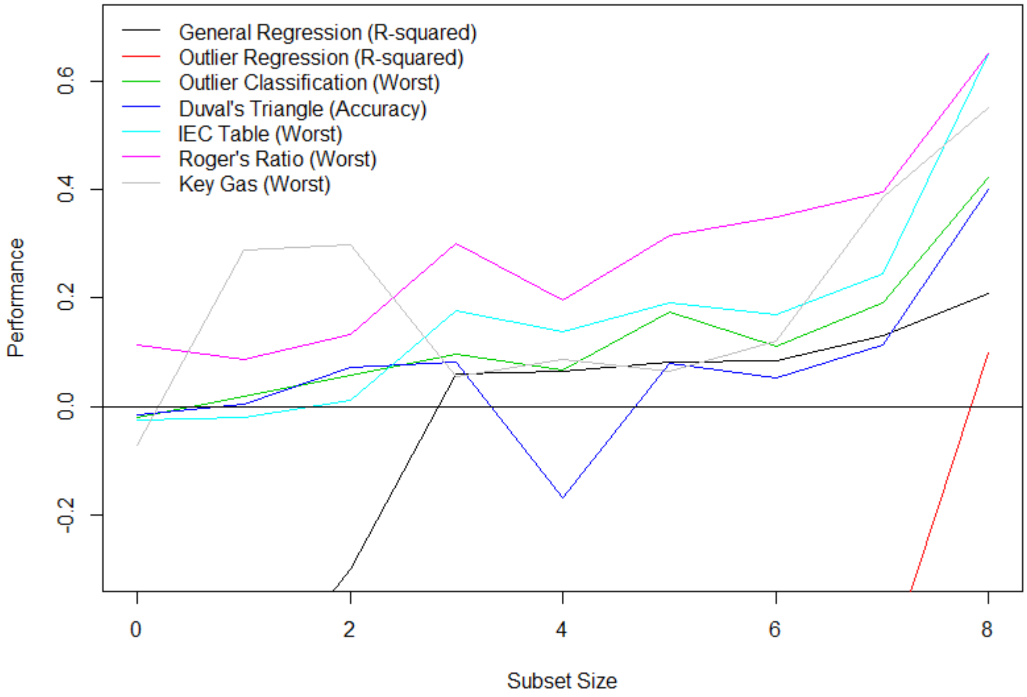


Figure 2. Plot for the full dataset of the various metrics compared with their thresholds.

6. CONCLUSION AND FUTURE WORK

The objectives set were ambitious and innovative, and this work was, to the best of our knowledge, the first to tackle them. Our most important contribution entails the identification of DGA subsets of size 3, which coupled with our ML methodology should allow any DGA based model to obtain comparable results with a fraction of the usual gases.

Furthermore, even without the ML methods developed, the two identified sets (H2, C2H2, CH4) and (H2, C2H2, C2H4), which are of equal size to those used by Duval's triangle and smaller than those of other traditional methods, while permitting potentially improved performance, can be used as a stepping stone for further research, as these combinations are not used by any existing method.

Finally, two novel data imputation techniques were developed, RSI for numerical variables and CSI for categorical ones, obtaining better results than the tested alternatives.

6.1. Limitations

Although great results were obtained, this work still presents some limitations. The acquired dataset is responsible for a majority having insufficient size, while being distributed over many very different transformers; large amounts of missing values and no information of real failures.

The remaining set of limitations is computational in nature. A very simple subset selection algorithm had to be used due to time constraints, while many, less but still important, preprocessing and modeling variations could not be tested. Furthermore, several model hyperparameter configurations were discarded for being too computationally expensive, mostly for the MLP models.

6.2. Future Work

The majority of the intended future work aims at surpassing the aforementioned limitations. The acquisition of a more complete data set, testing of different configurations and the identification of more efficient implementations of the methods used would all be essential to improve the results. Particular emphasis is placed on gathering and employing real fault data to validate our approach.

Another interesting path of research involves the further study of the developed RSI and CSI techniques, comparing these with a richer set of imputation methods and in a larger number of varied datasets and tasks.

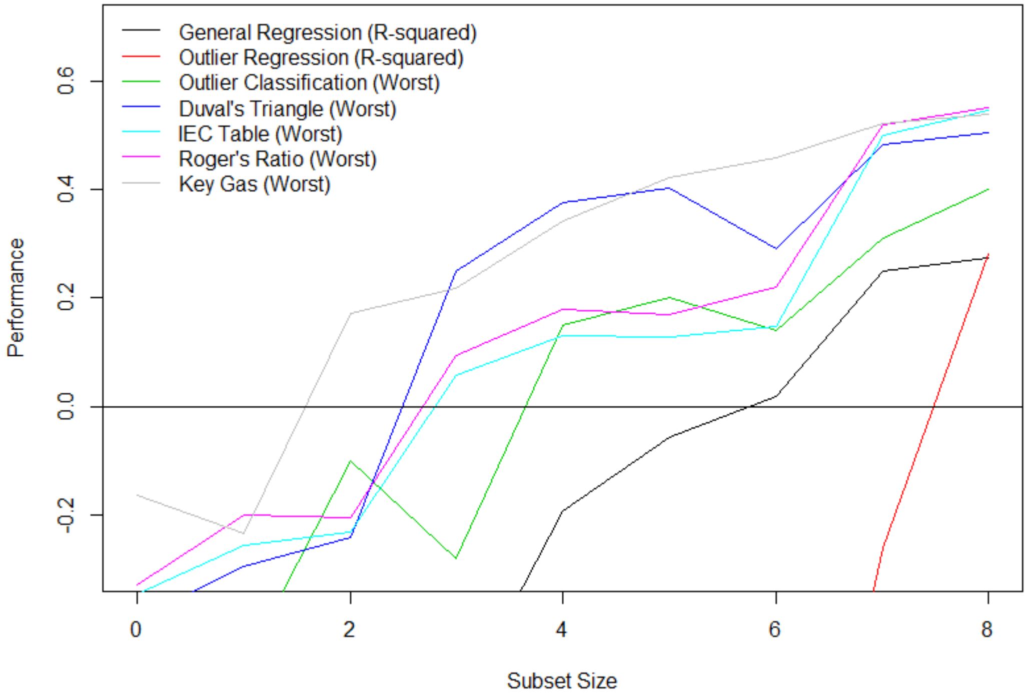


Figure 3. Plot for the DGA only dataset of the various metrics compared with their thresholds.

ACKNOWLEDGMENT

This work was supported by Fundo Europeu de Desenvolvimento Regional (FEDER) and Programa Operacional Competitividade e Internacionalização through the project TRF4p0 - TRANSFORMER4.0: DIGITAL REVOLUTION OF POWER TRANSFORMERS with reference POCI-01-0247-FEDER-045926.

REFERENCES

Andrianto, Afandi, A. N., Aripriharta, Fadlika, I., Putro, S. C., & Fajariawan, A. H. (2020). Analysis of maintenance scheduling transformer oil using Markov method. In *Aip conference proceedings* (Vol. 2228, p. 030020).

Chauhan, R. S., & Sinha, A. G. (2015). *Internal fault detection in three phase transformer using machine learning methods* (Unpublished doctoral dissertation).

Dhonge, D. D., Swamin, P., & Thosar, A. (2015). 'Developing Artificial Neural Network (ANN) Model for Fault Diagnosis of Power Transformer Using Dissolved Gas Analysis (DGA). *International Journal of Scientific & Engineering Research*, *6*(7).

Duval, M. (1989). Dissolved gas analysis: It can save your transformer. *IEEE Electrical Insulation Magazine*, *5*(6), 22–27.

Gavrilovs, G., & Vītoliņa, S. (2011). Identification of Power Transformer's Failure and Risk Source. In *Proceedings of the 52st annual international scientific conference,(october)* (pp. 1–4).

Georgilakis, P. S., Katsigiannis, J. A., Valavanis, K. P., & Souflaris, A. T. (2006). A systematic stochastic petri net based methodology for transformer fault diagnosis and repair actions. *Journal of Intelligent and Robotic Systems*, *45*(2), 181–201.

Liew, A. W.-C., Law, N.-F., & Yan, H. (2011). Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Briefings in bioinformatics*, *12*(5), 498–513.

Londo, L., Çelo, M., & Bualoti, R. (2015). Assessment of Transformer Condition Using the Improve Key Gas Methods. *International Journal of Engineering Research & Technology (IJERT)*, *4*, 48–55.

Mao, P. (2000). *Power transformer fault diagnosis based on wavelet transform and artificial neural network* (Unpublished doctoral dissertation). University of Bath.

Miranda, V., & Castro, A. R. G. (2005). Improving the IEC table for transformer failure diagnosis with knowledge extraction from neural networks. *IEEE transactions on power delivery*, *20*(4), 2509–2516.

Mirowski, P., & LeCun, Y. (2012). Statistical machine learning and dissolved gas analysis: a review. *IEEE Transactions on Power Delivery*, *27*(4), 1791–1799.

Naresh, R., Sharma, V., & Vashisth, M. (2008). An integrated neural fuzzy approach for fault diagnosis of transformers. *IEEE transactions on power delivery*, *23*(4), 2017–2024.

Nurmanova, V., Akhmetov, Y., Bagheri, M., Zollanvari, A., Gharehpetian, G. B., & Phung, T. (2020). A New Transformer FRA Test Setup for Advanced Interpretation and Winding Short-circuit Prediction. In *2020 ieee international conference on environment and electrical engineering and 2020 ieee industrial and commercial power systems europe (eeeic/i&cps europe)* (pp. 1–6).

Ravi, N. N., Drus, S. M., & Krishnan, P. S. (2019). Data mining techniques for transformer failure prediction model: A systematic literature review. In *2019 ieee 9th symposium on computer applications & industrial electronics (iscaie)* (pp. 305–309).

Sarajcev, P., Jakus, D., & Vasilj, J. (2020). Optimal scheduling of power transformers preventive maintenance with Bayesian statistical learning and influence diagramslabel. *Journal of Cleaner Production*, 120850.

Silva-Ramírez, E.-L., Pino-Mejías, R., López-Coello, M., & Cubiles-de-la Vega, M.-D. (2011). Missing value imputation on missing completely at random data using multilayer perceptrons. *Neural Networks*, *24*(1), 121–129.

Sun, H.-C., Huang, Y.-C., & Huang, C.-M. (2012). Fault diagnosis of power transformers using computational intelligence: A review. *Energy Procedia*, *14*, 1226–1231.

Tan, F., Fu, X., Zhang, Y., & Bourgeois, A. G. (2008). A genetic algorithm-based method for feature subset selection. *Soft Computing*, *12*(2), 111–120.

Velásquez, R. M. A., Lara, J. V. M., & Melgar, A. (2019). Converting data into knowledge for preventing failures in power transformers. *Engineering Failure Analysis*, *101*, 215–229.

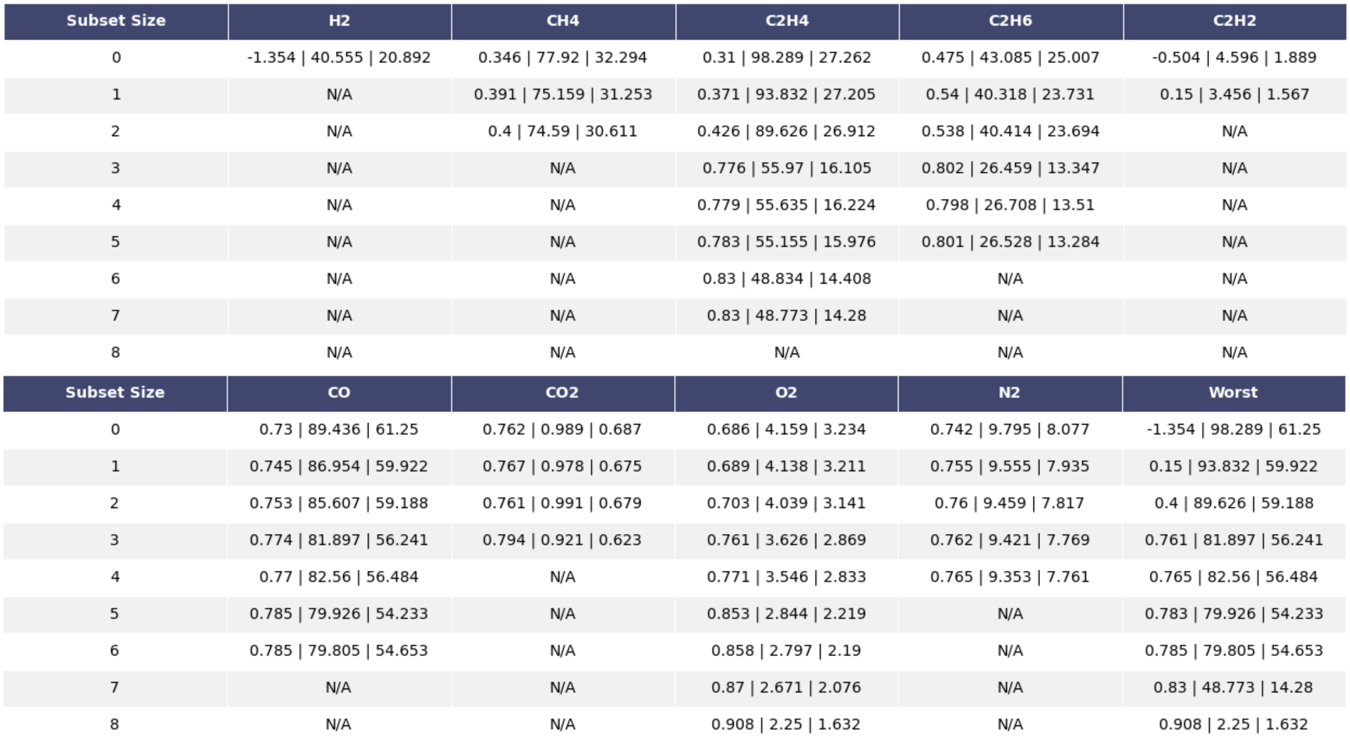
Ward, S. A. (2003). Evaluating transformer condition using dga oil analysis. In *2003 annual report conference on electrical insulation and dielectric phenomena* (pp. 463–468).

Zhao, Z., Tang, C., Zhou, Q., Xu, L., Gui, Y., & Yao, C. (2017). Identification of power transformer winding mechanical fault types based on online IFRA by support vector machine. *Energies*, *10*(12), 2022.

APPENDIX

Due to the large quantity and size of the tables referenced in Section 5, which would disrupt the structure of the document, we moved them to the Appendix.

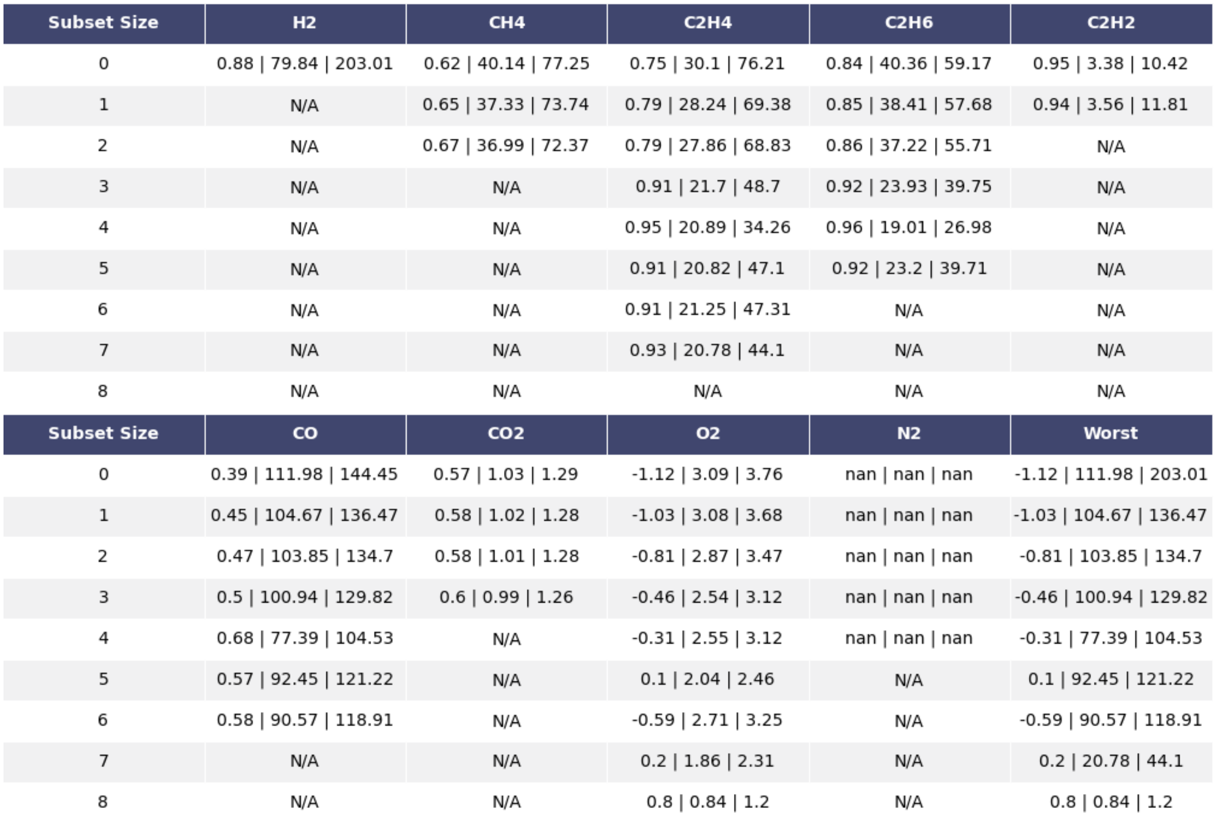
Table 1. Table of the Full dataset general regression metrics. Results in the form R-squared | RMSE | MAE



Subset Size	H2	CH4	C2H4	C2H6	C2H2
0	-1.354 | 40.555 | 20.892	0.346 | 77.92 | 32.294	0.31 | 98.289 | 27.262	0.475 | 43.085 | 25.007	-0.504 | 4.596 | 1.889
1	N/A	0.391 | 75.159 | 31.253	0.371 | 93.832 | 27.205	0.54 | 40.318 | 23.731	0.15 | 3.456 | 1.567
2	N/A	0.4 | 74.59 | 30.611	0.426 | 89.626 | 26.912	0.538 | 40.414 | 23.694	N/A
3	N/A	N/A	0.776 | 55.97 | 16.105	0.802 | 26.459 | 13.347	N/A
4	N/A	N/A	0.779 | 55.635 | 16.224	0.798 | 26.708 | 13.51	N/A
5	N/A	N/A	0.783 | 55.155 | 15.976	0.801 | 26.528 | 13.284	N/A
6	N/A	N/A	0.83 | 48.834 | 14.408	N/A	N/A
7	N/A	N/A	0.83 | 48.773 | 14.28	N/A	N/A
8	N/A	N/A	N/A	N/A	N/A

Subset Size	CO	CO2	O2	N2	Worst
0	0.73 | 89.436 | 61.25	0.762 | 0.989 | 0.687	0.686 | 4.159 | 3.234	0.742 | 9.795 | 8.077	-1.354 | 98.289 | 61.25
1	0.745 | 86.954 | 59.922	0.767 | 0.978 | 0.675	0.689 | 4.138 | 3.211	0.755 | 9.555 | 7.935	0.15 | 93.832 | 59.922
2	0.753 | 85.607 | 59.188	0.761 | 0.991 | 0.679	0.703 | 4.039 | 3.141	0.76 | 9.459 | 7.817	0.4 | 89.626 | 59.188
3	0.774 | 81.897 | 56.241	0.794 | 0.921 | 0.623	0.761 | 3.626 | 2.869	0.762 | 9.421 | 7.769	0.761 | 81.897 | 56.241
4	0.77 | 82.56 | 56.484	N/A	0.771 | 3.546 | 2.833	0.765 | 9.353 | 7.761	0.765 | 82.56 | 56.484
5	0.785 | 79.926 | 54.233	N/A	0.853 | 2.844 | 2.219	N/A	0.783 | 79.926 | 54.233
6	0.785 | 79.805 | 54.653	N/A	0.858 | 2.797 | 2.19	N/A	0.785 | 79.805 | 54.653
7	N/A	N/A	0.87 | 2.671 | 2.076	N/A	0.83 | 48.773 | 14.28
8	N/A	N/A	0.908 | 2.25 | 1.632	N/A	0.908 | 2.25 | 1.632

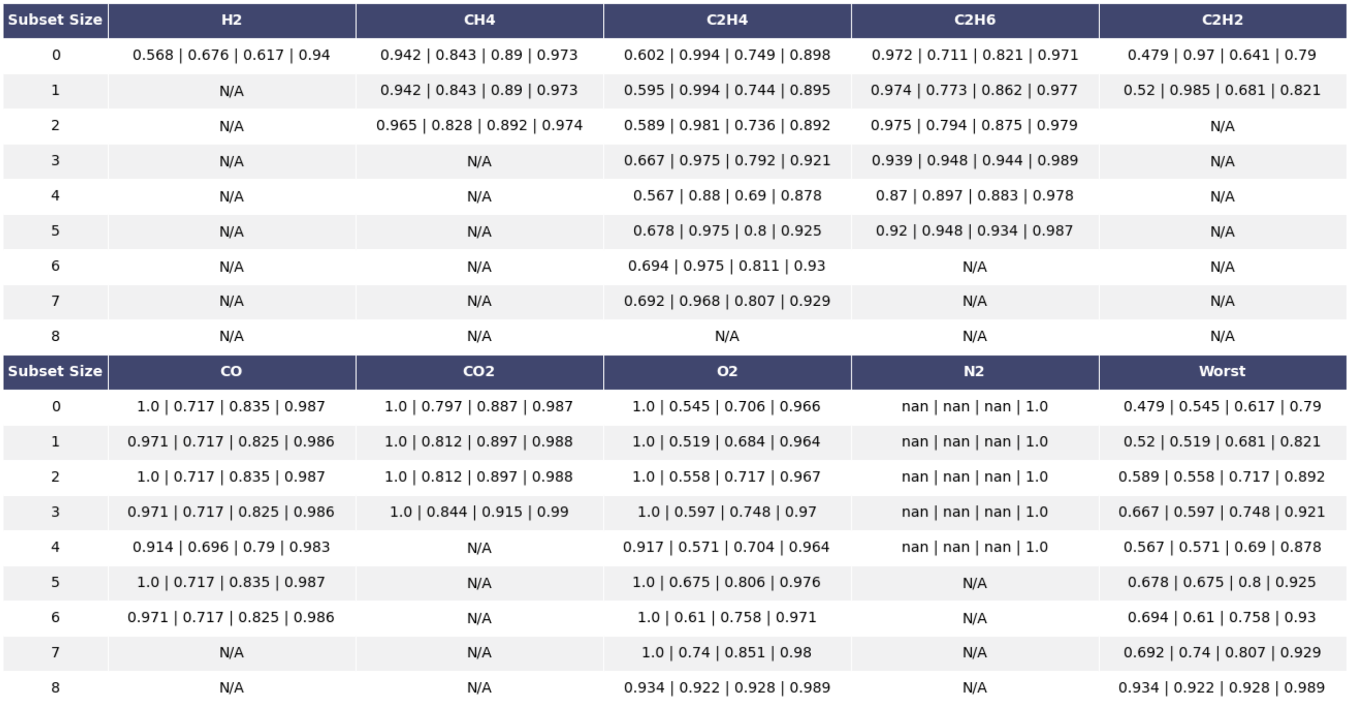
Table 2. Table of the Full dataset outlier regression metrics. Results in the form R-squared | RMSE | MAE



Subset Size	H2	CH4	C2H4	C2H6	C2H2
0	0.88 | 79.84 | 203.01	0.62 | 40.14 | 77.25	0.75 | 30.1 | 76.21	0.84 | 40.36 | 59.17	0.95 | 3.38 | 10.42
1	N/A	0.65 | 37.33 | 73.74	0.79 | 28.24 | 69.38	0.85 | 38.41 | 57.68	0.94 | 3.56 | 11.81
2	N/A	0.67 | 36.99 | 72.37	0.79 | 27.86 | 68.83	0.86 | 37.22 | 55.71	N/A
3	N/A	N/A	0.91 | 21.7 | 48.7	0.92 | 23.93 | 39.75	N/A
4	N/A	N/A	0.95 | 20.89 | 34.26	0.96 | 19.01 | 26.98	N/A
5	N/A	N/A	0.91 | 20.82 | 47.1	0.92 | 23.2 | 39.71	N/A
6	N/A	N/A	0.91 | 21.25 | 47.31	N/A	N/A
7	N/A	N/A	0.93 | 20.78 | 44.1	N/A	N/A
8	N/A	N/A	N/A	N/A	N/A

Subset Size	CO	CO2	O2	N2	Worst
0	0.39 | 111.98 | 144.45	0.57 | 1.03 | 1.29	-1.12 | 3.09 | 3.76	nan | nan | nan	-1.12 | 111.98 | 203.01
1	0.45 | 104.67 | 136.47	0.58 | 1.02 | 1.28	-1.03 | 3.08 | 3.68	nan | nan | nan	-1.03 | 104.67 | 136.47
2	0.47 | 103.85 | 134.7	0.58 | 1.01 | 1.28	-0.81 | 2.87 | 3.47	nan | nan | nan	-0.81 | 103.85 | 134.7
3	0.5 | 100.94 | 129.82	0.6 | 0.99 | 1.26	-0.46 | 2.54 | 3.12	nan | nan | nan	-0.46 | 100.94 | 129.82
4	0.68 | 77.39 | 104.53	N/A	-0.31 | 2.55 | 3.12	nan | nan | nan	-0.31 | 77.39 | 104.53
5	0.57 | 92.45 | 121.22	N/A	0.1 | 2.04 | 2.46	N/A	0.1 | 92.45 | 121.22
6	0.58 | 90.57 | 118.91	N/A	-0.59 | 2.71 | 3.25	N/A	-0.59 | 90.57 | 118.91
7	N/A	N/A	0.2 | 1.86 | 2.31	N/A	0.2 | 20.78 | 44.1
8	N/A	N/A	0.8 | 0.84 | 1.2	N/A	0.8 | 0.84 | 1.2

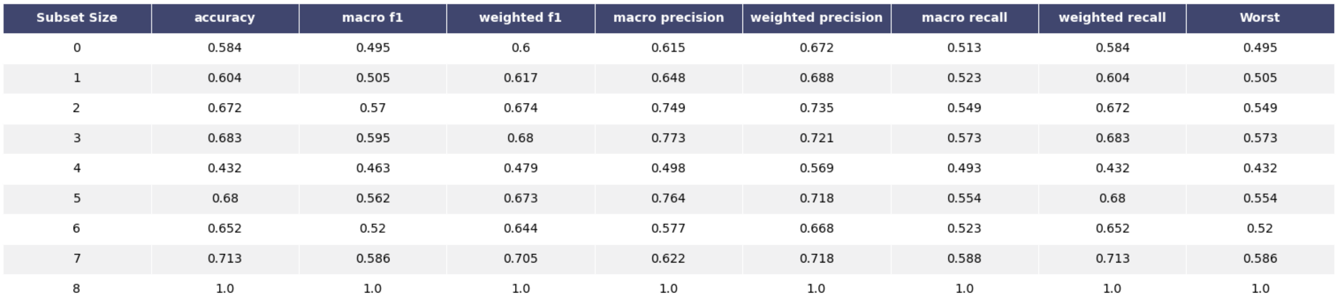
Table 3. Table of the Full dataset outlier binary classification metrics. Results in the form Precision | Recall | F1 | Accuracy.



Subset Size	H2	CH4	C2H4	C2H6	C2H2
0	0.568 | 0.676 | 0.617 | 0.94	0.942 | 0.843 | 0.89 | 0.973	0.602 | 0.994 | 0.749 | 0.898	0.972 | 0.711 | 0.821 | 0.971	0.479 | 0.97 | 0.641 | 0.79
1	N/A	0.942 | 0.843 | 0.89 | 0.973	0.595 | 0.994 | 0.744 | 0.895	0.974 | 0.773 | 0.862 | 0.977	0.52 | 0.985 | 0.681 | 0.821
2	N/A	0.965 | 0.828 | 0.892 | 0.974	0.589 | 0.981 | 0.736 | 0.892	0.975 | 0.794 | 0.875 | 0.979	N/A
3	N/A	N/A	0.667 | 0.975 | 0.792 | 0.921	0.939 | 0.948 | 0.944 | 0.989	N/A
4	N/A	N/A	0.567 | 0.88 | 0.69 | 0.878	0.87 | 0.897 | 0.883 | 0.978	N/A
5	N/A	N/A	0.678 | 0.975 | 0.8 | 0.925	0.92 | 0.948 | 0.934 | 0.987	N/A
6	N/A	N/A	0.694 | 0.975 | 0.811 | 0.93	N/A	N/A
7	N/A	N/A	0.692 | 0.968 | 0.807 | 0.929	N/A	N/A
8	N/A	N/A	N/A	N/A	N/A

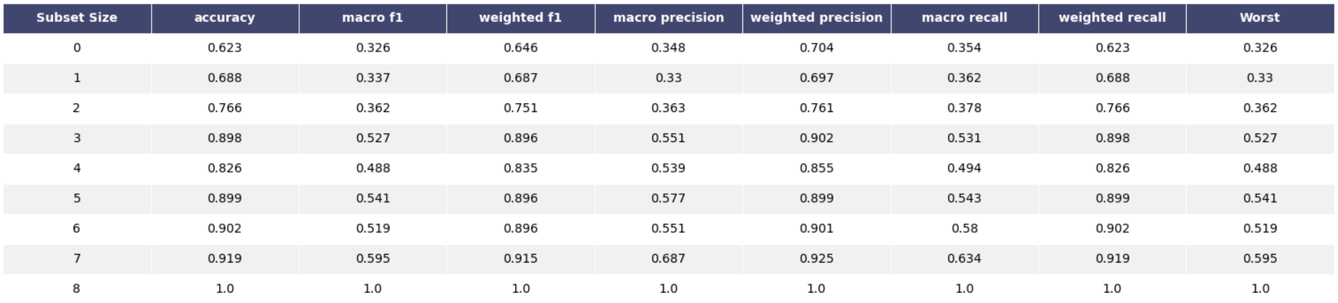
Subset Size	CO	CO2	O2	N2	Worst
0	1.0 | 0.717 | 0.835 | 0.987	1.0 | 0.797 | 0.887 | 0.987	1.0 | 0.545 | 0.706 | 0.966	nan | nan | nan | 1.0	0.479 | 0.545 | 0.617 | 0.79
1	0.971 | 0.717 | 0.825 | 0.986	1.0 | 0.812 | 0.897 | 0.988	1.0 | 0.519 | 0.684 | 0.964	nan | nan | nan | 1.0	0.52 | 0.519 | 0.681 | 0.821
2	1.0 | 0.717 | 0.835 | 0.987	1.0 | 0.812 | 0.897 | 0.988	1.0 | 0.558 | 0.717 | 0.967	nan | nan | nan | 1.0	0.589 | 0.558 | 0.717 | 0.892
3	0.971 | 0.717 | 0.825 | 0.986	1.0 | 0.844 | 0.915 | 0.99	1.0 | 0.597 | 0.748 | 0.97	nan | nan | nan | 1.0	0.667 | 0.597 | 0.748 | 0.921
4	0.914 | 0.696 | 0.79 | 0.983	N/A	0.917 | 0.571 | 0.704 | 0.964	nan | nan | nan | 1.0	0.567 | 0.571 | 0.69 | 0.878
5	1.0 | 0.717 | 0.835 | 0.987	N/A	1.0 | 0.675 | 0.806 | 0.976	N/A	0.678 | 0.675 | 0.8 | 0.925
6	0.971 | 0.717 | 0.825 | 0.986	N/A	1.0 | 0.61 | 0.758 | 0.971	N/A	0.694 | 0.61 | 0.758 | 0.93
7	N/A	N/A	1.0 | 0.74 | 0.851 | 0.98	N/A	0.692 | 0.74 | 0.807 | 0.929
8	N/A	N/A	0.934 | 0.922 | 0.928 | 0.989	N/A	0.934 | 0.922 | 0.928 | 0.989

Table 4. Table of the full dataset Duval's triangle classification metrics.



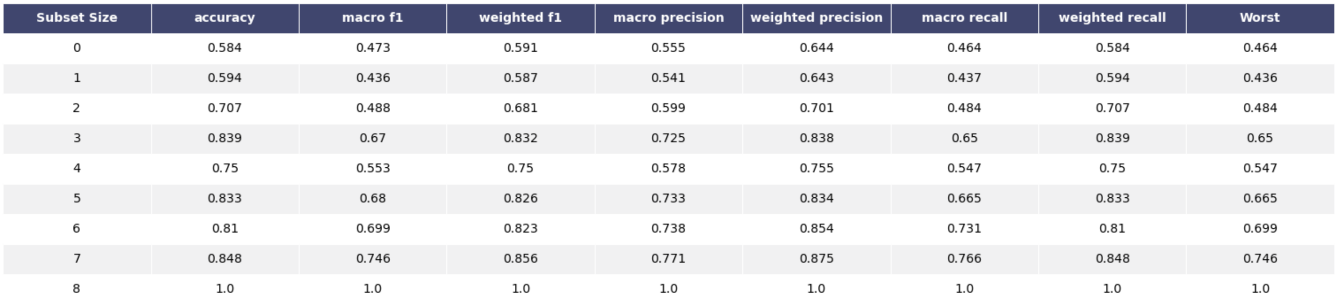
Subset Size	accuracy	macro f1	weighted f1	macro precision	weighted precision	macro recall	weighted recall	Worst
0	0.584	0.495	0.6	0.615	0.672	0.513	0.584	0.495
1	0.604	0.505	0.617	0.648	0.688	0.523	0.604	0.505
2	0.672	0.57	0.674	0.749	0.735	0.549	0.672	0.549
3	0.683	0.595	0.68	0.773	0.721	0.573	0.683	0.573
4	0.432	0.463	0.479	0.498	0.569	0.493	0.432	0.432
5	0.68	0.562	0.673	0.764	0.718	0.554	0.68	0.554
6	0.652	0.52	0.644	0.577	0.668	0.523	0.652	0.52
7	0.713	0.586	0.705	0.622	0.718	0.588	0.713	0.586
8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Table 5. Table of the full dataset IEC table classification metrics.



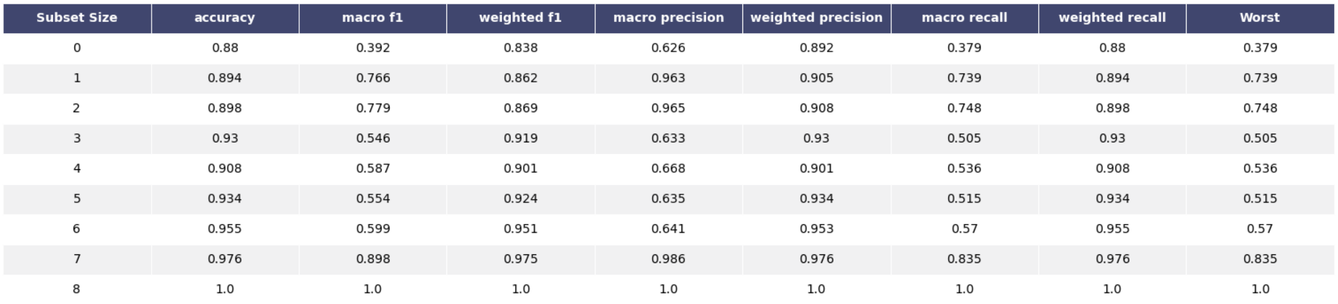
Subset Size	accuracy	macro f1	weighted f1	macro precision	weighted precision	macro recall	weighted recall	Worst
0	0.623	0.326	0.646	0.348	0.704	0.354	0.623	0.326
1	0.688	0.337	0.687	0.33	0.697	0.362	0.688	0.33
2	0.766	0.362	0.751	0.363	0.761	0.378	0.766	0.362
3	0.898	0.527	0.896	0.551	0.902	0.531	0.898	0.527
4	0.826	0.488	0.835	0.539	0.855	0.494	0.826	0.488
5	0.899	0.541	0.896	0.577	0.899	0.543	0.899	0.541
6	0.902	0.519	0.896	0.551	0.901	0.58	0.902	0.519
7	0.919	0.595	0.915	0.687	0.925	0.634	0.919	0.595
8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Table 6. Table of the full dataset Rogers ratio classification metrics.



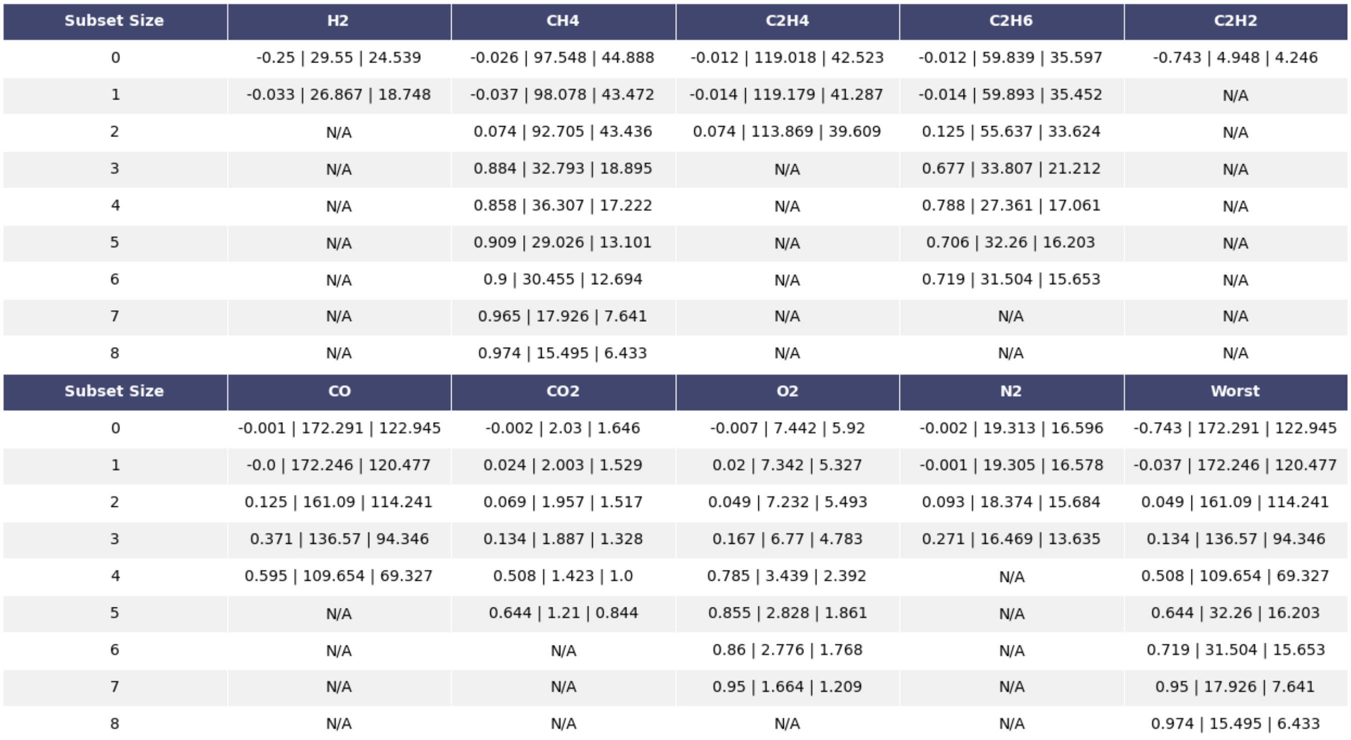
Subset Size	accuracy	macro f1	weighted f1	macro precision	weighted precision	macro recall	weighted recall	Worst
0	0.584	0.473	0.591	0.555	0.644	0.464	0.584	0.464
1	0.594	0.436	0.587	0.541	0.643	0.437	0.594	0.436
2	0.707	0.488	0.681	0.599	0.701	0.484	0.707	0.484
3	0.839	0.67	0.832	0.725	0.838	0.65	0.839	0.65
4	0.75	0.553	0.75	0.578	0.755	0.547	0.75	0.547
5	0.833	0.68	0.826	0.733	0.834	0.665	0.833	0.665
6	0.81	0.699	0.823	0.738	0.854	0.731	0.81	0.699
7	0.848	0.746	0.856	0.771	0.875	0.766	0.848	0.746
8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Table 7. Table of the full dataset key gas method classification metrics.



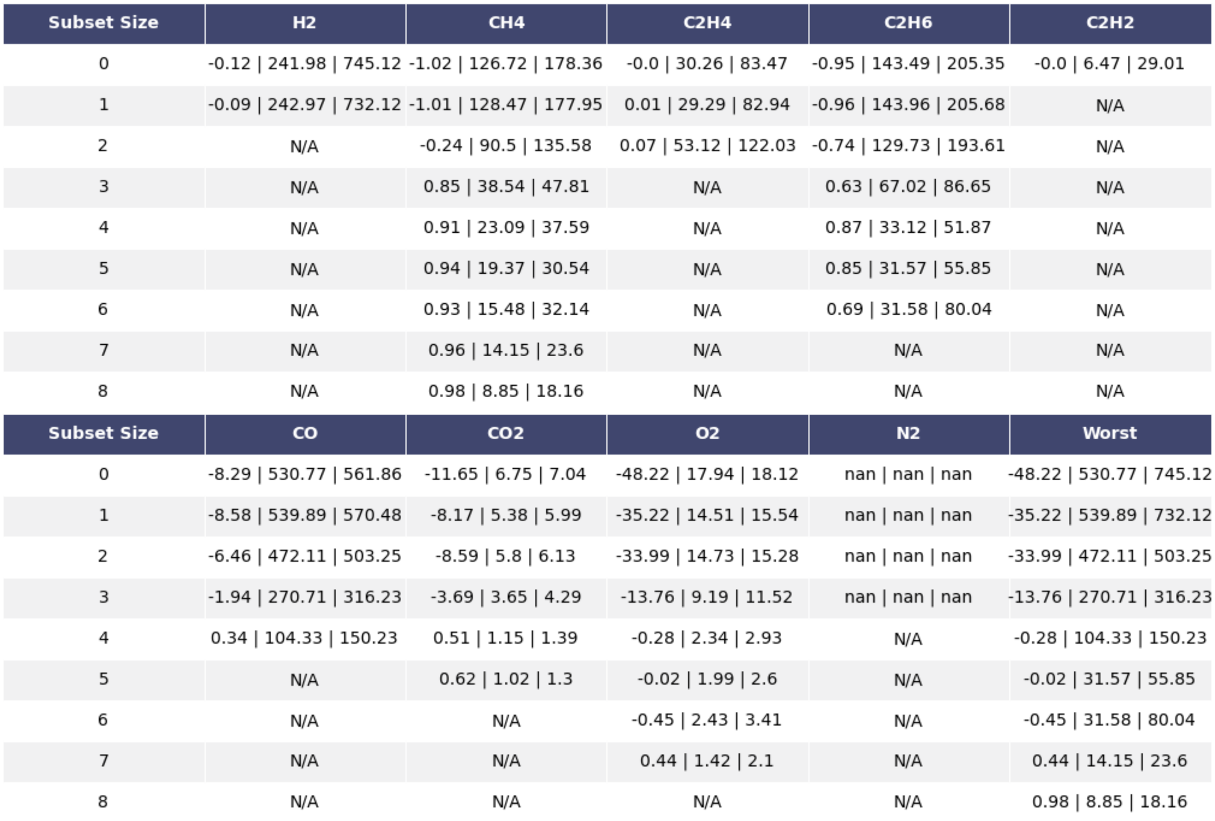
Subset Size	accuracy	macro f1	weighted f1	macro precision	weighted precision	macro recall	weighted recall	Worst
0	0.88	0.392	0.838	0.626	0.892	0.379	0.88	0.379
1	0.894	0.766	0.862	0.963	0.905	0.739	0.894	0.739
2	0.898	0.779	0.869	0.965	0.908	0.748	0.898	0.748
3	0.93	0.546	0.919	0.633	0.93	0.505	0.93	0.505
4	0.908	0.587	0.901	0.668	0.901	0.536	0.908	0.536
5	0.934	0.554	0.924	0.635	0.934	0.515	0.934	0.515
6	0.955	0.599	0.951	0.641	0.953	0.57	0.955	0.57
7	0.976	0.898	0.975	0.986	0.976	0.835	0.976	0.835
8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Table 8. Table of the DGA only dataset general regression metrics. Results in the form R-squared — RMSE — MAE



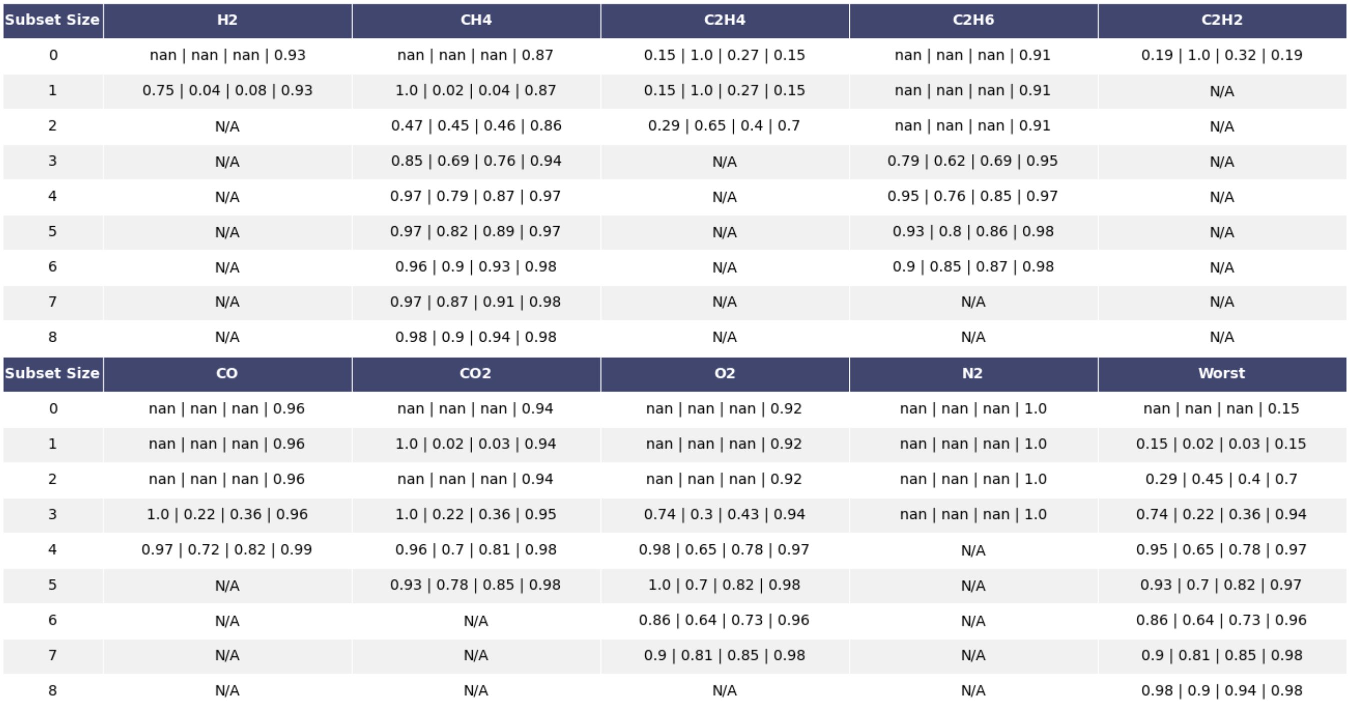
Subset Size	H2	CH4	C2H4	C2H6	C2H2	CO	CO2	O2	N2	Worst
0	-0.25 | 29.55 | 24.539	-0.026 | 97.548 | 44.888	-0.012 | 119.018 | 42.523	-0.012 | 59.839 | 35.597	-0.743 | 4.948 | 4.246	-0.001 | 172.291 | 122.945	-0.002 | 2.03 | 1.646	-0.007 | 7.442 | 5.92	-0.002 | 19.313 | 16.596	-0.743 | 172.291 | 122.945
1	-0.033 | 26.867 | 18.748	-0.037 | 98.078 | 43.472	-0.014 | 119.179 | 41.287	-0.014 | 59.893 | 35.452	N/A	-0.0 | 172.246 | 120.477	0.024 | 2.003 | 1.529	0.02 | 7.342 | 5.327	-0.001 | 19.305 | 16.578	-0.037 | 172.246 | 120.477
2	N/A	0.074 | 92.705 | 43.436	0.074 | 113.869 | 39.609	0.125 | 55.637 | 33.624	N/A	0.125 | 161.09 | 114.241	0.069 | 1.957 | 1.517	0.049 | 7.232 | 5.493	0.093 | 18.374 | 15.684	0.049 | 161.09 | 114.241
3	N/A	0.884 | 32.793 | 18.895	N/A	0.677 | 33.807 | 21.212	N/A	0.371 | 136.57 | 94.346	0.134 | 1.887 | 1.328	0.167 | 6.77 | 4.783	0.271 | 16.469 | 13.635	0.134 | 136.57 | 94.346
4	N/A	0.858 | 36.307 | 17.222	N/A	0.788 | 27.361 | 17.061	N/A	0.595 | 109.654 | 69.327	0.508 | 1.423 | 1.0	0.785 | 3.439 | 2.392	N/A	0.508 | 109.654 | 69.327
5	N/A	0.909 | 29.026 | 13.101	N/A	0.706 | 32.26 | 16.203	N/A	N/A	0.644 | 1.21 | 0.844	0.855 | 2.828 | 1.861	N/A	0.644 | 32.26 | 16.203
6	N/A	0.9 | 30.455 | 12.694	N/A	0.719 | 31.504 | 15.653	N/A	N/A	N/A	0.86 | 2.776 | 1.768	N/A	0.719 | 31.504 | 15.653
7	N/A	0.965 | 17.926 | 7.641	N/A	N/A	N/A	N/A	N/A	0.95 | 1.664 | 1.209	N/A	0.95 | 17.926 | 7.641
8	N/A	0.974 | 15.495 | 6.433	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.974 | 15.495 | 6.433

Table 9. Table of the DGA only dataset outlier regression metrics. Results in the form R-squared — RMSE — MAE



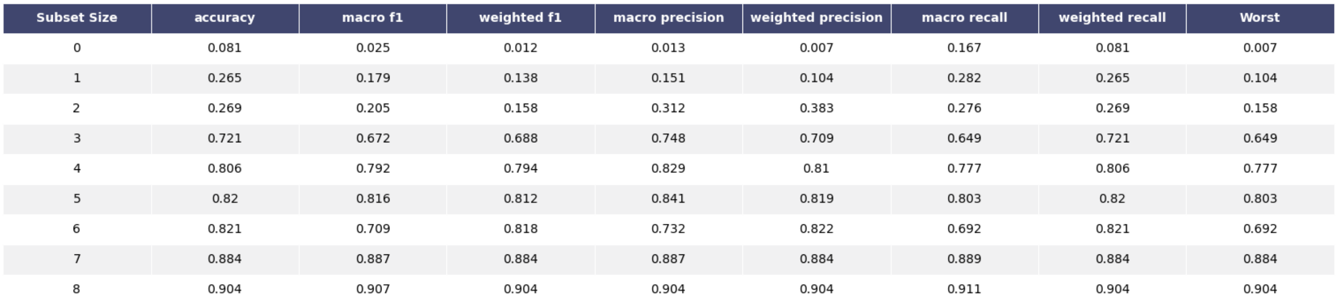
<table>
<tr><th>Subset Size</th><th>H2</th><th>CH4</th><th>C2H4</th><th>C2H6</th><th>C2H2</th></tr>
<tr><td>0</td><td>-0.12 | 241.98 | 745.12</td><td>-1.02 | 126.72 | 178.36</td><td>-0.0 | 30.26 | 83.47</td><td>-0.95 | 143.49 | 205.35</td><td>-0.0 | 6.47 | 29.01</td></tr>
<tr><td>1</td><td>-0.09 | 242.97 | 732.12</td><td>-1.01 | 128.47 | 177.95</td><td>0.01 | 29.29 | 82.94</td><td>-0.96 | 143.96 | 205.68</td><td>N/A</td></tr>
<tr><td>2</td><td>N/A</td><td>-0.24 | 90.5 | 135.58</td><td>0.07 | 53.12 | 122.03</td><td>-0.74 | 129.73 | 193.61</td><td>N/A</td></tr>
<tr><td>3</td><td>N/A</td><td>0.85 | 38.54 | 47.81</td><td>N/A</td><td>0.63 | 67.02 | 86.65</td><td>N/A</td></tr>
<tr><td>4</td><td>N/A</td><td>0.91 | 23.09 | 37.59</td><td>N/A</td><td>0.87 | 33.12 | 51.87</td><td>N/A</td></tr>
<tr><td>5</td><td>N/A</td><td>0.94 | 19.37 | 30.54</td><td>N/A</td><td>0.85 | 31.57 | 55.85</td><td>N/A</td></tr>
<tr><td>6</td><td>N/A</td><td>0.93 | 15.48 | 32.14</td><td>N/A</td><td>0.69 | 31.58 | 80.04</td><td>N/A</td></tr>
<tr><td>7</td><td>N/A</td><td>0.96 | 14.15 | 23.6</td><td>N/A</td><td>N/A</td><td>N/A</td></tr>
<tr><td>8</td><td>N/A</td><td>0.98 | 8.85 | 18.16</td><td>N/A</td><td>N/A</td><td>N/A</td></tr>
<tr><th>Subset Size</th><th>CO</th><th>CO2</th><th>O2</th><th>N2</th><th>Worst</th></tr>
<tr><td>0</td><td>-8.29 | 530.77 | 561.86</td><td>-11.65 | 6.75 | 7.04</td><td>-48.22 | 17.94 | 18.12</td><td>nan | nan | nan</td><td>-48.22 | 530.77 | 745.12</td></tr>
<tr><td>1</td><td>-8.58 | 539.89 | 570.48</td><td>-8.17 | 5.38 | 5.99</td><td>-35.22 | 14.51 | 15.54</td><td>nan | nan | nan</td><td>-35.22 | 539.89 | 732.12</td></tr>
<tr><td>2</td><td>-6.46 | 472.11 | 503.25</td><td>-8.59 | 5.8 | 6.13</td><td>-33.99 | 14.73 | 15.28</td><td>nan | nan | nan</td><td>-33.99 | 472.11 | 503.25</td></tr>
<tr><td>3</td><td>-1.94 | 270.71 | 316.23</td><td>-3.69 | 3.65 | 4.29</td><td>-13.76 | 9.19 | 11.52</td><td>nan | nan | nan</td><td>-13.76 | 270.71 | 316.23</td></tr>
<tr><td>4</td><td>0.34 | 104.33 | 150.23</td><td>0.51 | 1.15 | 1.39</td><td>-0.28 | 2.34 | 2.93</td><td>N/A</td><td>-0.28 | 104.33 | 150.23</td></tr>
<tr><td>5</td><td>N/A</td><td>0.62 | 1.02 | 1.3</td><td>-0.02 | 1.99 | 2.6</td><td>N/A</td><td>-0.02 | 31.57 | 55.85</td></tr>
<tr><td>6</td><td>N/A</td><td>N/A</td><td>-0.45 | 2.43 | 3.41</td><td>N/A</td><td>-0.45 | 31.58 | 80.04</td></tr>
<tr><td>7</td><td>N/A</td><td>N/A</td><td>0.44 | 1.42 | 2.1</td><td>N/A</td><td>0.44 | 14.15 | 23.6</td></tr>
<tr><td>8</td><td>N/A</td><td>N/A</td><td>N/A</td><td>N/A</td><td>0.98 | 8.85 | 18.16</td></tr>
</table>

Table 10. Table of the DGA only dataset outlier binary classification metrics. Results in the form Precision | Recall | F1 | Accuracy



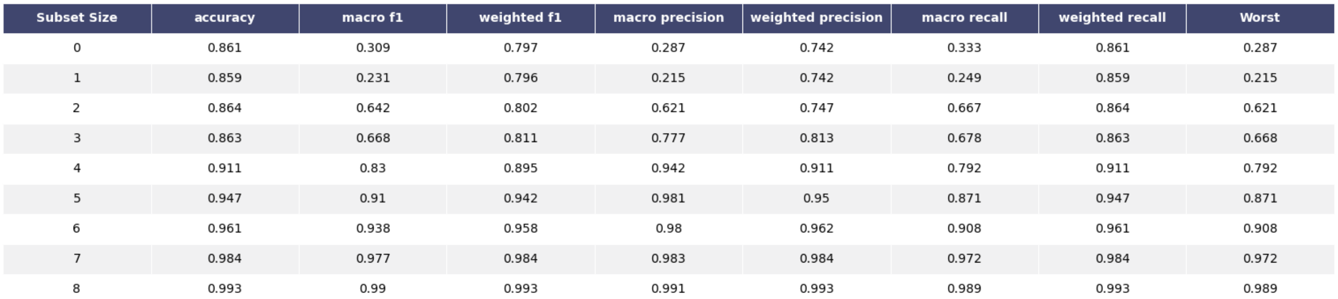
<table>
<tr><th>Subset Size</th><th>H2</th><th>CH4</th><th>C2H4</th><th>C2H6</th><th>C2H2</th></tr>
<tr><td>0</td><td>nan | nan | nan | 0.93</td><td>nan | nan | nan | 0.87</td><td>0.15 | 1.0 | 0.27 | 0.15</td><td>nan | nan | nan | 0.91</td><td>0.19 | 1.0 | 0.32 | 0.19</td></tr>
<tr><td>1</td><td>0.75 | 0.04 | 0.08 | 0.93</td><td>1.0 | 0.02 | 0.04 | 0.87</td><td>0.15 | 1.0 | 0.27 | 0.15</td><td>nan | nan | nan | 0.91</td><td>N/A</td></tr>
<tr><td>2</td><td>N/A</td><td>0.47 | 0.45 | 0.46 | 0.86</td><td>0.29 | 0.65 | 0.4 | 0.7</td><td>nan | nan | nan | 0.91</td><td>N/A</td></tr>
<tr><td>3</td><td>N/A</td><td>0.85 | 0.69 | 0.76 | 0.94</td><td>N/A</td><td>0.79 | 0.62 | 0.69 | 0.95</td><td>N/A</td></tr>
<tr><td>4</td><td>N/A</td><td>0.97 | 0.79 | 0.87 | 0.97</td><td>N/A</td><td>0.95 | 0.76 | 0.85 | 0.97</td><td>N/A</td></tr>
<tr><td>5</td><td>N/A</td><td>0.97 | 0.82 | 0.89 | 0.97</td><td>N/A</td><td>0.93 | 0.8 | 0.86 | 0.98</td><td>N/A</td></tr>
<tr><td>6</td><td>N/A</td><td>0.96 | 0.9 | 0.93 | 0.98</td><td>N/A</td><td>0.9 | 0.85 | 0.87 | 0.98</td><td>N/A</td></tr>
<tr><td>7</td><td>N/A</td><td>0.97 | 0.87 | 0.91 | 0.98</td><td>N/A</td><td>N/A</td><td>N/A</td></tr>
<tr><td>8</td><td>N/A</td><td>0.98 | 0.9 | 0.94 | 0.98</td><td>N/A</td><td>N/A</td><td>N/A</td></tr>
<tr><th>Subset Size</th><th>CO</th><th>CO2</th><th>O2</th><th>N2</th><th>Worst</th></tr>
<tr><td>0</td><td>nan | nan | nan | 0.96</td><td>nan | nan | nan | 0.94</td><td>nan | nan | nan | 0.92</td><td>nan | nan | nan | 1.0</td><td>nan | nan | nan | 0.15</td></tr>
<tr><td>1</td><td>nan | nan | nan | 0.96</td><td>1.0 | 0.02 | 0.03 | 0.94</td><td>nan | nan | nan | 0.92</td><td>nan | nan | nan | 1.0</td><td>0.15 | 0.02 | 0.03 | 0.15</td></tr>
<tr><td>2</td><td>nan | nan | nan | 0.96</td><td>nan | nan | nan | 0.94</td><td>nan | nan | nan | 0.92</td><td>nan | nan | nan | 1.0</td><td>0.29 | 0.45 | 0.4 | 0.7</td></tr>
<tr><td>3</td><td>1.0 | 0.22 | 0.36 | 0.96</td><td>1.0 | 0.22 | 0.36 | 0.95</td><td>0.74 | 0.3 | 0.43 | 0.94</td><td>nan | nan | nan | 1.0</td><td>0.74 | 0.22 | 0.36 | 0.94</td></tr>
<tr><td>4</td><td>0.97 | 0.72 | 0.82 | 0.99</td><td>0.96 | 0.7 | 0.81 | 0.98</td><td>0.98 | 0.65 | 0.78 | 0.97</td><td>N/A</td><td>0.95 | 0.65 | 0.78 | 0.97</td></tr>
<tr><td>5</td><td>N/A</td><td>0.93 | 0.78 | 0.85 | 0.98</td><td>1.0 | 0.7 | 0.82 | 0.98</td><td>N/A</td><td>0.93 | 0.7 | 0.82 | 0.97</td></tr>
<tr><td>6</td><td>N/A</td><td>N/A</td><td>0.86 | 0.64 | 0.73 | 0.96</td><td>N/A</td><td>0.86 | 0.64 | 0.73 | 0.96</td></tr>
<tr><td>7</td><td>N/A</td><td>N/A</td><td>0.9 | 0.81 | 0.85 | 0.98</td><td>N/A</td><td>0.9 | 0.81 | 0.85 | 0.98</td></tr>
<tr><td>8</td><td>N/A</td><td>N/A</td><td>N/A</td><td>N/A</td><td>0.98 | 0.9 | 0.94 | 0.98</td></tr>
</table>

Table 11. Table of the DGA only Duval's triangle classification metrics.



Subset Size	accuracy	macro f1	weighted f1	macro precision	weighted precision	macro recall	weighted recall	Worst
0	0.081	0.025	0.012	0.013	0.007	0.167	0.081	0.007
1	0.265	0.179	0.138	0.151	0.104	0.282	0.265	0.104
2	0.269	0.205	0.158	0.312	0.383	0.276	0.269	0.158
3	0.721	0.672	0.688	0.748	0.709	0.649	0.721	0.649
4	0.806	0.792	0.794	0.829	0.81	0.777	0.806	0.777
5	0.82	0.816	0.812	0.841	0.819	0.803	0.82	0.803
6	0.821	0.709	0.818	0.732	0.822	0.692	0.821	0.692
7	0.884	0.887	0.884	0.887	0.884	0.889	0.884	0.884
8	0.904	0.907	0.904	0.904	0.904	0.911	0.904	0.904

Table 12. Table of the DGA only Key Gas classification metrics.



Subset Size	accuracy	macro f1	weighted f1	macro precision	weighted precision	macro recall	weighted recall	Worst
0	0.861	0.309	0.797	0.287	0.742	0.333	0.861	0.287
1	0.859	0.231	0.796	0.215	0.742	0.249	0.859	0.215
2	0.864	0.642	0.802	0.621	0.747	0.667	0.864	0.621
3	0.863	0.668	0.811	0.777	0.813	0.678	0.863	0.668
4	0.911	0.83	0.895	0.942	0.911	0.792	0.911	0.792
5	0.947	0.91	0.942	0.981	0.95	0.871	0.947	0.871
6	0.961	0.938	0.958	0.98	0.962	0.908	0.961	0.908
7	0.984	0.977	0.984	0.983	0.984	0.972	0.984	0.972
8	0.993	0.99	0.993	0.991	0.993	0.989	0.993	0.989