Hybrid AI-Subject Matter Expert Solution for Evaluating the Health Index of Oil Distribution Transformers

Augustin Cathignol[1], Victor Thuillie-Demont[2], Ludovica Baldi[3], Laurent Micheau[4], Jean-Pierre Petitpretre[5], Amelle Ouberehil[6]

[1,2,3] *Schneider Electric, 160 Avenue des Martyrs, 38000 Grenoble, France*
augustin.cathignol@se.com, victor.thuilliedemont@se.com, ludovica.baldi@se.com

[4,5,6] *Transfo Services, Rue Jacques Lieutaud, 13200, France*
laurent.micheau@se.com, jean-pierre.petitpretre@se.com, amelle.ouberehil@se.com

ABSTRACT

Reliability of oil distribution transformers is paramount, ensuring a stable flow of electricity and shielding from potential fire hazards. The internal insulation system of these transformers utilizes a combination of oil and paper. As the oil circulates through the active part of the system, it collects gaseous and physical traces of existing or past defects or degradations, providing a holistic view of the transformer's health, and allowing for early detection of problems and predictive maintenance. While various and mainly data-driven methods have been developed to calculate a transformer health index from oil samples, they lack accuracy due to limited data. This paper proposes a novel hybrid approach that leverages both Artificial Intelligence and Subject Matter Expertise to enhance the health estimation of oil distribution transformers. Our methodology utilizes a substantial dataset exceeding 65,600 analyzed oil samples, coupled with the valuable knowledge of domain experts. This combined approach achieves an accuracy exceeding 95%, suitable for real-world industrial applications. Furthermore, we introduce a risk management feature that strengthens the ability to identify transformers at high risk of failure. Notably, the health index estimation is implemented as a semi-automatic process, retaining the "expert in the loop" principle for managing critical and ambiguous cases.

1. INTRODUCTION AND PROBLEM STATEMENT

Distribution transformers convert high-voltage electricity from transmission lines into usable power for homes, businesses, and industries. Their reliability is paramount, ensuring a stable and continuous flow of electricity, shielding us from power outages, and potential fire hazards. Regular inspections, advanced fault detection systems, condition monitoring, and proper maintenance are crucial for those transformers. The internal insulation system of a transformer is provided by both oil and paper. As the oil circulates through the active part of the system, it collects gaseous and physical traces of existing or past defects or degradations. Therefore, it provides a holistic view of the transformer's health, allowing for early detection of problems and enabling predictive maintenance. In such a process, oil samples are extracted and analyzed in the laboratory regarding their physical and chemical properties (dielectric strength, acidity, humidity, color, and dissolved gas concentration). Sample extraction and analysis are done on a regular basis that can range from a several months to year periodicity. The results are interpreted by experienced subject matter experts who attribute a Health Index (HI) to the transformer and guide maintenance actions that could be required. The huge number of oil distribution transformers currently in operation and the limited number of experienced subject matter experts available to estimate the HI of these devices, motivates to support them. Several methods were developed to automatically compute the HI. A review of HI automatic assessment techniques for distribution and power transformers was proposed by Quynh T. Tran, Kevin Davies, Leon Roose, Puthawat Wiriyakitikun, Jaktupong Janjampop, Eleonora Riva Sanseverino and Gaetano Zizzo (2020). Some of the techniques rely on on-line data, which is not the scope of this study. Most of the techniques that rely on off-line data are fully data driven. The HI estimations done by experts not only rely on standardized combinatory calculations, but also reflect the human expertise in interpreting results. Consequently, it is difficult to translate them into mathematical formulas and several studies implemented fuzzy logic approaches, as proposed by Ahmed E. B. Abu-Elanien M.M.A. Salama, and M. Ibrahim (2012). Whatever the data-driven algorithm used, these methods are poorly explainable, which reduces their capacity to be adopted in real operations. Only some of them rely on hybrid approach combining expert knowledge, either for feature selection, as proposed by Khalil Ibrahim, R.M. Sharkawy, H.K. Temraz and M.M.A. Salama (2016), or for uncertainty management, using Bayesian modeling, as proposed by P. Sarajcev, D. Jakus, J. Vasilj, M. Nikolic (2018). Finally, while these approaches offer intriguing avenues, they rely on very limited data set, with fewer than 100 samples (Ahmed E. B. et al. (2012), P. Sarajcev et al (2018), Atefeh Dehghani Ashkezari et al. (2012), Jahanzaib Javid et al. (2021), Ahmed E. B. Abu-Elanien et al. (2011)). These restricted datasets are unlikely to encompass the exhaustive spectrum of parameter combinations, rendering them poorly fit for real-world implementation, especially considering the safety concerns.

Now, the motivation of this work is precisely to build a prognostics and health management solution that is suitable for industrial usage, meaning performant enough, resilient enough, and preserving safety in any case. The core of this work is an original hybrid approach relying on both Artificial Intelligence (AI) and Subject Matter Expertise (SME), making the most of AI and human expertise. Practically, this approach is supported by more than 60 000 oil samples that were analyzed and from which experts provided Health Index estimations. Indeed, the idea is to train a Machine Learning (ML) algorithm to estimate HI from oil samples analyses results (Figure 1). The performance criterion that is pursued is the global accuracy of the health estimation, with a special focus on the ability of detecting transformers at risk, for obvious security reasons. The solution is also expected to be explainable and to be suitable with an "keeping expert in the loop" approach. Indeed, the target solution is not a fully automated solution, but rather a mostly automated solution that will keep experts in the loop for managing the most ambiguous and critical cases.

The paper describes the health estimation global solution, starting with the data science steps, from data collection, data cleaning, outliers' detection, imputation for managing missing data, up to the model selection and validation. It also emphasizes the way subject matter expertise was combined with AI techniques. It describes the risk management method that was introduced to minimize the risk of failing to detect at risk transformers. Finally, it practically describes the global health evaluation process in an industrial context with a "keeping expert in the loop" approach that was mentioned above.

2. Hybrid Health estimation method

2.1. Introduction

The method that is used is said hybrid approach as it is based on both machine learning and subject matter expertise. The machine learning pillar is a standard approach from a data science point of view, supported by subject matter expertise at all stages (feature engineering, outliers' detection, missing value management, result validation). The expertise is also explicitly integrated in a rules-based approach that complements the machine learning approach to refine health estimation. The current section covers the following technical data science steps: data collection, data cleaning, outliers' detection and validation, missing values management, models benchmarking, including oversampling and/or subsampling methods, and rule-based classification.

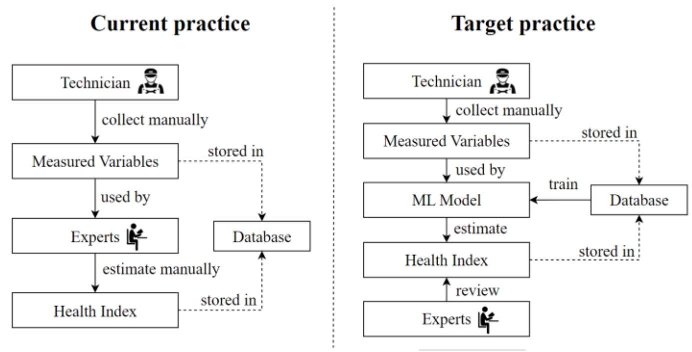

Figure 1: Current and target practice.

2.2. Data collection

This study focuses on distribution transformers which power is less than 3150 kVA and using mineral oil. Data from oil analyses over the past 10 years were used, representing approximately 65,600 analyses for 40,000 distinct transformers (as some of them have been analyzed several times all along their lifecycle). The predictive variables identified in the data are the levels of dissolved gases, color, acidity and humidity, as the dielectric strength. The target of our study is the Health Index (HI), which was estimated by experienced subject matter experts, based on the oil analysis result. Predictive variables and target are given in Figure 2.

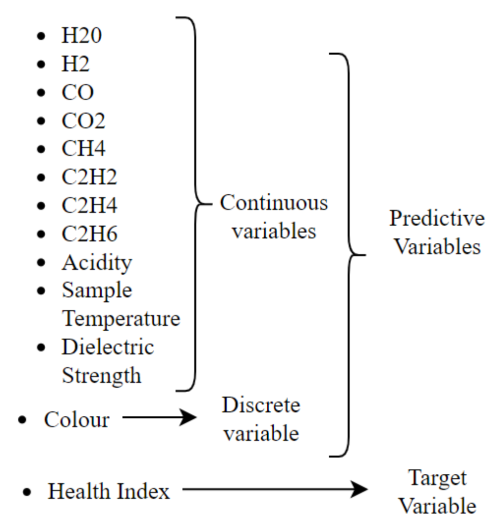

Figure 2: Predictive variables and target.

The distribution of the Health Index is intrinsically highly unbalanced: for most of the analyses the Health Index is evaluated by experts at 1 (around 90%), followed by a smaller number of analyses with a Health Index of 2 (around 9%), and finally an even smaller number of analyses with a Health Index of 3 (around 1%). A Health Index equal to 1 means that the transformer is perfectly healthy, a Health Index of 2 is an intermediate status showing some anomalies, non-serious at this stage but requiring some surveillance, and a Health Index of 3 means critical anomalies that require immediate maintenance actions.

2.3. Outliers' detection

The data cover more than 10 years of oil analyses carried out by technical experts. As the analysis process is manual, it may be possible to have some errors in the data. In this study potential errors were tracked using an anomaly detection approach. Then, each anomaly was reviewed, confirmed or not to be an error, by technical experts. The anomaly detection used in this study relies on a statistical test: the Hotelling's T-squared test. Hotelling's test is a multivariate statistical test used to determine if there are significant differences between the means of two groups in a multivariate space. In anomaly detection, it allows us to compare the mean vector and covariance matrix of a single data point (or a small group of data points) to those of a reference group. If the data point falls outside a certain threshold based on the T-squared statistic, it's flagged as an anomaly. When using Hotelling's T-squared test, it's important to ensure that the data follow a multivariate normal distribution, as it is an assumption it relies on. In the present case, dissolved gas concentrations follow an exponential distribution, so a logarithmic transformation was needed before running the test. Using a 99% confidence level for setting the T-squared threshold, 125 analyses (out of a total of 65,600 analyses, meaning 0.2%) were identified with potential errors. As atypical does not mean error, it was necessary to have those analyses reviewed by technical experts to confirm whether or not the presence of errors. In the end only 17 analyses were confirmed with errors and removed from the dataset. For the other analyses initially flagged the level of some dissolved gases was exceptionally high but perfectly credible.

2.4. Models benchmarking

To choose the machine learning algorithm that will be the basis of our hybrid approach, we perform a model screening, by doing a classification test with nine different algorithms, on all analyses for which no data are missing. The imputation method for managing missing data is studied later, after finetuning the hyperparameters of each algorithm. According to the results shown in Table 1 and obtained with a 12,000 analyses validation dataset, it appears that the most performant algorithm is the Random Forest Classifier. Accuracy is not the only performance criterion that we want to meet. Indeed, interpreting the model and validating, to some level, its consistency with the subject matter expert's way of working is another relevant criterion. A Random Forest algorithm doesn't allow to clearly identify the reasons behind a given prediction, as the final output is a combination of many decision trees, making it difficult to pinpoint the logic for each prediction. However, Random Forest algorithms usually embed feature importance techniques that show how much a specific feature contributes to the overall mode. Such a technique was used, and it could be verified that the top 3 most influencing features, Hydrogen (H_2), dielectric strength and acetylene (C_2H_2), are consistent with the subject matter expertise, which historically allows to estimate the HI. Indeed, Hydrogen (H_2) is the gas produced by most technical faults, so an analysis of oil sample done by experts always starts with this gas. The dielectric strength (also called rigidity) provide experts with a good indication of the water present in the transformer oil and therefore of the risk during operation: a too low dielectric strength induces a risk of flashover. Even in the field, this parameter is checked after certain maintenance operations, before restarting the equipment. Finally, acetylene (C_2H_2) is synonymous for experts with an electric arc, and therefore a major electrical fault. This consistency between algorithm feature importance and subject matter expertise gives trust in health estimation algorithm.

Table 1: Classifiers' performances obtained on the validation dataset.

Classifier	Accuracy	AUC	Recall	Precision
Random Forest	0.96	0.93	0.96	0.95
Extreme Gradient Boosting	0.96	0.92	0.96	0.95
Light Gradient Boosting	0.96	0.93	0.96	0.95
Gradient Boosting	0.96	0.93	0.96	0.95
Extra Tree	0.95	0.92	0.95	0.95
Ada Boost	0.95	0.85	0.95	0.95
Logistic Regression	0.94	0.82	0.94	0.93
Decision Tree	0.94	0.79	0.94	0.94
K Neighbors	0.93	0.67	0.93	0.90

2.5. Imputation of missing values

In the dataset used for the study (representing 65,600 analyses), the data corresponding to the dissolved gas concentrations are complete for all analyses (no missing value for any of the dissolved gases). Concerning the other data, a value is missing for less than 10% of the analyses. Different approaches to impute missing values were tested. The goal of this step is to make the most of all analyses, even those for which a value is missing. First, the simplest imputation method was chosen as a reference: imputation by the mean. This method consists in replacing the missing values for a given feature by the average value of the feature itself. Then, the data imputation was tested using two other methods that follow a similar approach: iterative imputation. Rather than simply replacing missing values with point estimates, iterative imputation makes multiple passes over the data, using the observed values to estimate and fill in the missing values, and then repeating this process several times to improve the estimates. This allows for data variability and relationships between variables, providing more robust estimates. Iterative imputation methods may include statistical models or machine learning techniques to estimate missing values. wo variants were tested. The first one is a standard version of iterative imputation, that includes a linear regression to estimate missing values. This first variant is promising as there are significant correlations between some of the variables, as can be seen in Figure 3. Indeed, for each analysis, each input parameter can be quite well estimated thanks to the others. The second variant, also called Miss Forest, uses Random Forest models to predict and fill in missing data. It builds separate models for variables with missing data, using the available data to make accurate predictions. This method is effective for handling both continuous and categorical variables.

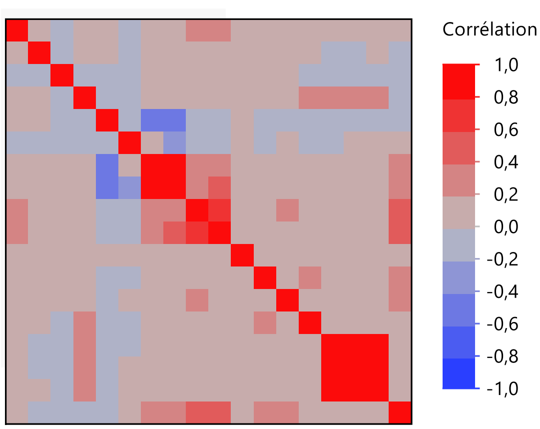

Figure 3: Correlation between parameters. Each row and each column are input variables (features).

Here we compare the output results of the Random Forest classifier using the three different imputers, on the analyses for which at least one value is missing. They all lead to the same accuracy, and according to the results shown in Table 2, it appears that the iterative imputation using a linear regression shows the best results in terms of precision for HI=1 class, recall for HI=3 class and proportion the falsely predicted HI 1 instead of 3. Therefore, the classic iterative imputer was chosen as imputation method.

Table 2: Imputation using iterative imputer, simple (mean) imputation, and Miss Forest.

Imputation method	Precision HI=1	Recall HI=3	Proportion of HI 3 diagnosed as HI 1
Iterative Imputer (Linear Regression)	**0.97**	**0.63**	**0.17**
Simple Imputation (Mean)	0.96	0.57	0.17
Miss Forest	0.96	0.57	0.38

2.6. Subject matter expert – Rule based Classification

2.6.1. Compliance with normative values

After having apply technical rules that must be respected to remain in compliance with the normative values, the expert predicted a Health Index. Concretely, these rules allow to frame the result. It is thus impossible in our context to highlight for a given analysis whose values would have exceeded the thresholds set by the normative standards.

2.6.2. Rules based on evolution through time

In most cases, a single analysis allows to set the HI of a transformer, but in some ambiguous cases, experts do use the past analyses of the same transformer (up to two additional past analyses) to refine their diagnostic. By combining last analysis results with the evolution in the dissolved gas concentrations between successive analyses, experts set the final HI. Such an approach was mimicked in the study to even improve the classification accuracy obtained from the last analyses, as shown in previous section. It led to a one-point increase in global accuracy.

2.7. Global scheme of the health estimation process

The global health estimation process, presented in Figure 4, is semi-automatic as the expert remains present in the process, to analyze and recommend maintenance actions for transformers whose Health Index is evaluated at 3, the most critical level. It is also hybrid because it relies on a machine learning core and on rules provided by the experts in the field.

It consists of three blocks:

- The machine learning prediction, based on the last analysis.
- The legal rules that ensure the compliance of any parameter of this last analysis.
- The expert rules that consider the evolution in dissolved gas concentration evolution through time, based on previous analyses.

The details of the process that includes those three main blocks are provided in Figure 4. In this figure, note that details about the cost matrix are provided in section 3.

Finally, once HI has been estimated, if it appears to be equal or superior to 2, an expert is asked to review the analyses and to provide recommendations in terms of maintenance and/or additional analyses to perform. It can also trigger a reclassification to class 1 (healthy transformer) if the expert concludes that a HI of 2 or 3 is not justified (Figure 5).

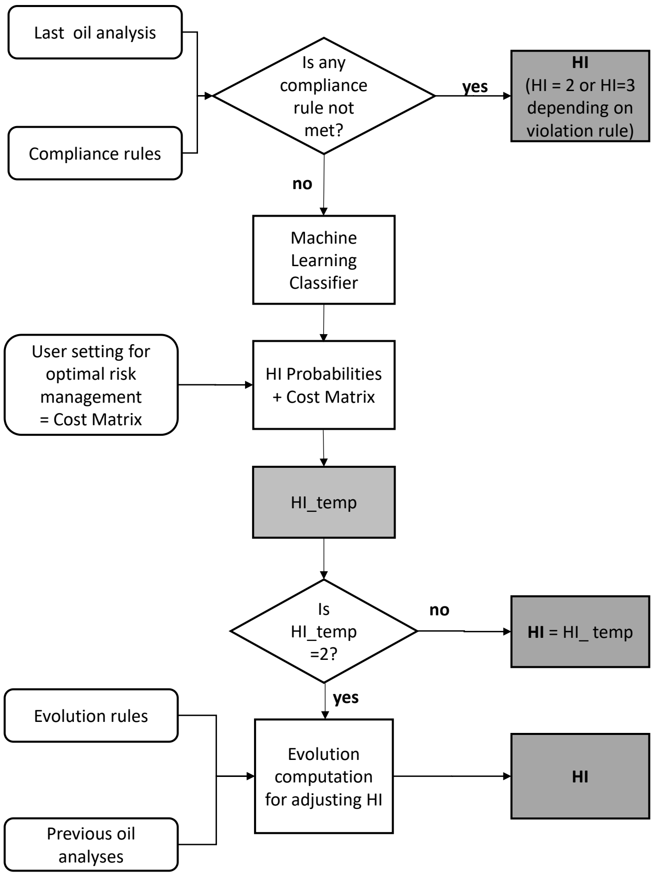

Figure 4: HI estimation synoptic. Inputs are on the left; computation are in the middle, and output is on the right.

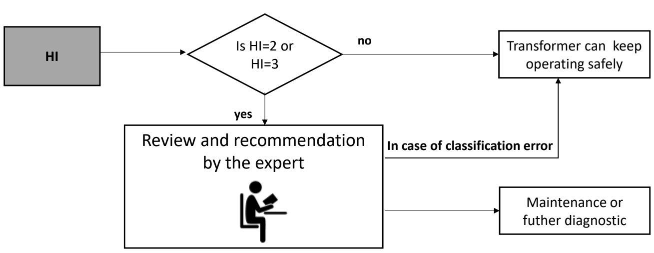

Figure 5: Health Index usage with "expert in the loop"

3. RISK MANAGEMENT FEATURE IN HEALTH ESTIMATION

3.1. Methodology

As described in the introduction, not detecting an at-risk transformer would lead to a risky situation that could have catastrophic consequences. On the other hand, wrongly categorizing a transformer as at-risk whereas it is healthy has minor consequences, as any at-risk transformer will be manually expertised. Indeed, the expert who would analyze such a transformer would set it back in the correct healthy category, with minor consequences, except the time spent for such analysis and action. This means that there is a different cost associated to false positive (healthy transformers wrongly detected as at-risk ones) and false negative (at risk transformers wrongly detected as healthy ones), with false negative being more penalizing than false positive. Also, such asymmetry depends on the context of usage. Indeed, having an at-risk transformer into the wild is always a situation to be avoided, but in some cases, it could be even more damaging than in other cases, considering the criticality of the systems that are supplied by the transformer (hospitals for instance) and considering the environment of the transformer and the risks in case of fire. In this context the goal is now to optimize the classification algorithm not regarding global accuracy, but to a cost function that considers various levels of false positive and false negative costs. Practically, we proceeded with the following steps:

- A cost matrix is defined, attributing some arbitrary weight to each of the errors, reflecting the higher cost of errors for false positive vs. false negative.
- From the classifier, the likelihood that the HI is 1, 2, or 3 is extracted thanks to the Random Forest that easily outputs a calibrated likelihood for each class.
- Considering the likelihood for each class and the cost of each possible choice for prediction, the HI is chosen so that it maximizes the total cost.

Figure 6 shows an example with three different settings (low, medium and high costs), with higher and higher cost for false negative. The goal of higher cost is to better detect at-risk transformers. Technically speaking, the goal is to increase the recall of at-risk transformers (HI=2 and, even more, HI=3). In this example, the likelihood of prediction of each HI is provided. Using a neutral cost matrix, the more likely HI would be selected. In this case HI is predicted as a 1. Using a medium cost matrix, because of the costs, HI is predicted as a 2 as it maximizes the global cost. Similarly, using a stronger cost matrix, named high, HI=3 is selected. The approach was generalized to 15 settings with increasing weights and tested on a 12,000 analyses validation dataset. This led to the results presented in Figure 7 and Figure 8, showing the expected effect on precision and recall: recall for HI=3 class increases as the setting gets more conservative (higher costs) and parallelly its precision decreases. Regarding HI=1 class, its precision increases and its recall decreases as the setting gets more conservative. As an intermediate class, HI=2 sees its precision decrease as the setting gets more conservative. Its recall first increases (as more analyses are correctly classified in HI=2 class, instead of HI=1 class) and then decreases: this because when the setting gets highly conservative the classification tends to incorrectly class HI=2 in the HI=3 class, as it can be seen in the confusion matrix (Figure 9).

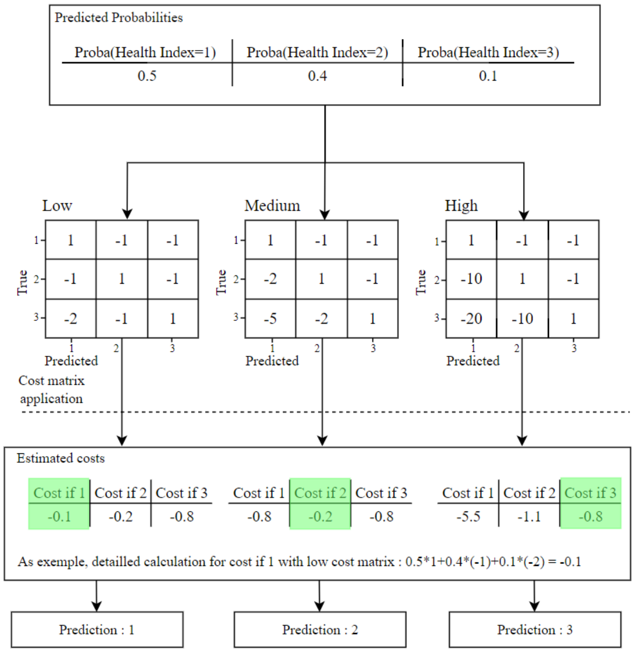

Figure 6: Illustration of cost matrix impact on prediction.

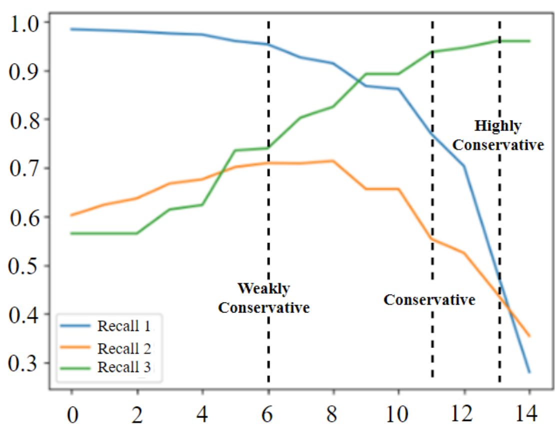

Figure 7: Recall obtained for HI equal to 1, 2 and with using increasingly settings, from 0 to 14.

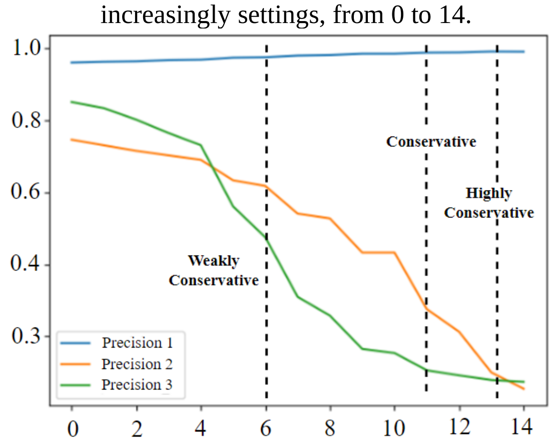

Figure 8: Precision obtained for HI equal to 1, 2 and 3 with using increasingly conservative settings, from 0 to 14.

In order to choose the best setting from the experts' point of view, we randomly selected 200 samples and provided the classification results for all the 15 settings, highlighting the correct results and the wrong ones (correct result means predicted HI equal to the HI estimated by the subject matter expert). This way, subject matter experts could easily see the conservatism level of each setting and choose the three settings that were the most pertinent to them, for covering the global range of criticality of the transformers' context. Those three settings, named weakly conservative, conservative and highly conservative cost matrices, are shown in Figure 7 and Figure 8.

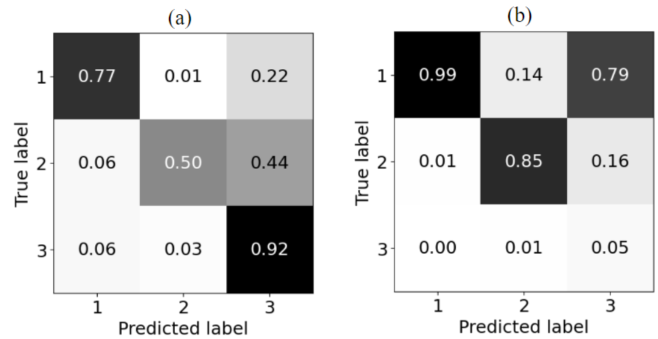

Figure 9: Confusion matrix normalized by rows and by columns for highly conservative cost matrix (respectively (a) and (b)). Results obtained on a 6k analyses dataset.

3.2. Results validation

The global algorithm (described in Figure 4) was tested on a new dataset of 6,000 analyses (called test dataset), using the three settings chosen by the subject matter experts. The results are shown in Table 3. They are consistent with the results obtained on the validation dataset. Particularly, precision for HI=1 class and recall for HI=3 class are very close. For instance, using the conservative cost matrix, precision for HI=1 class is 99% on the test dataset and 99% on the validation dataset, recall for HI=3 class is 95% on the validation dataset and 94% on the test dataset. Those performances and the risk management settings allow for industrial usage.

Table 3: Results obtained on the test dataset for weakly conservative, conservative, and highly conservative settings, focusing on precision for HI=1and recall for HI=3.

Total Number of Analyses=6000	Accuracy	Precision(HI=1)	Recall(HI=3)
Weakly Conservative Cost Matrix	**95%**	**0.98**	**0.75**
Conservative Cost Matrix	**91%**	**0.99**	**0.94**
Highly Conservative Cost Matrix	**76%**	**0.99**	**0.98**

4. CONCLUSION AND DISCUSSION

A semi-automatic, hybrid machine learning and expert-based approach for transformer maintenance has been developed. This approach is based on a very large number of analyses (65,600) carried out over more than 10 years and the technical experts who carried them out. It is called semi-automatic because the expert remains present in the process, especially to analyze and recommend maintenance actions for transformers whose Health Index is evaluated at 3, the most critical level. It is called hybrid because it relies on a machine learning core and on rules provided by the domain experts. The machine learning part goes beyond the application of combinatorial rules: it has captured the experience and practices of experts exposed to many analyses and their practical experience on many transformers, throughout their lifecycle, who are familiar with the signatures of faults and their probability of leading to more serious problems later. Like any machine learning algorithm, the performance of this solution relies on a big amount of data for training. Knowing that this solution is a hybrid solution that also relies on expertise, any industry with advanced expertise necessarily also possesses a large volume of data. Therefore, it is suitable to any industrial player in the field. As in most classification problems, it is not possible to simultaneously improve precision and recall, or in other words, minimize false alarms and minimize non-detections. This is why we have introduced a setting that allows us to prioritize one or the other, depending on the context of use.

Our overall health estimation system relies on:

- This machine learning-based estimation core.
- Legal and safety rules that need to be verified.

Calculations commonly used by experts based on the evolution of dissolved gas concentrations trough time to discriminate the most ambiguous cases.

- The expert who will confirm the critical cases (2 or 3) and provide an appropriate maintenance or further analyses recommendations.

This transformers health estimation enabling predictive maintenance is now deployed on the cloud as an API that is exposed to users whose use can also be done directly through a web application. Today, this process relies on discrete oil analyses; tomorrow, with more and more embedded monitoring in transformers, it is possible to perform real-time analyses. The hybrid approach can be preserved, but this time the machine learning core can rely on the time series and be even more sensitive to any degradation and more accurate in failure prediction.

Finally, it should be noted that this methodology is transferable to many other application domains beyond transformers.

REFERENCES

Quynh T. Tran, Kevin Davies, Leon Roose, Puthawat Wiriyakitikun, Jaktupong Janjampop, Eleonora Riva Sanseverino and Gaetano Zizzo (2020). A Review of Health Assessment Techniques for Distribution Transformers in Smart Distribution Grids, *Applied sciences, 2020, Vol.10 (22), p.8115*

Ahmed E. B. Abu-Elanien, M.M.A. Salama, M. Ibrahim (2012). Calculation of a Health Index for Oil-Immersed Transformers Rated Under 69 kV Using Fuzzy Logic. *IEEE transactions on power delivery, 2012, Vol.27 (4), p.2029-2036*

Khalil Ibrahim, R.M. Sharkawy, H.K. Temraz and M.M.A. Salama (2016). Selection criteria for oil transformer measurements to calculate the Health Index. *IEEE transactions on dielectrics and electrical insulation, 2016, Vol.23 (6), p.3397-340*

P. Sarajcev, D. Jakus, J. Vasilj, M. Nikolic (2018). Analysis of Transformer Health Index Using Bayesian Statistical Models. *2018 International Conference on Smart and Sustainable Technologies, 2018, p.1-7*

Atefeh Dehghani Ashkezari, Hui Ma, Chandima Ekanayake, Tapan K. Saha (2012). Multivariate analysis for correlations among different transformer oil parameters to determine transformer health index. *2012 IEEE Power and Energy Society General Meeting, 2012, p.1-7*

Jahanzaib Javid, Muhammad Ali Mughal, Mustansir Karim (2021). Using kNN Algorithm for classification of Distribution transformers Health index. *2021 International Conference on Innovative Computing (ICIC), 2021, p.1-6*

Ahmed E. B. Abu-Elanien, M. M. A. Salama, Malak Ibrahim (2011). Determination of transformer health condition using artificial neural networks, *2011 International Symposium on Innovations in Intelligent Systems and Applications, 2011, p.1-5*

BIOGRAPHIES

Augustin Cathignol is with Schneider Electric. As both a principal Data Scientist and expert in Reliability, monitoring,

failure prediction, he is the domain leader for Prognostics and Artificial Intelligence solutions developed by the Artificial Intelligence Hub, for Schneider electric assets and systems. Working closely with subject matter experts, he believes in hybrid solutions that make the most of AI and physics-based models provided by the experts. Prior to joining Schneider Electric, he was for 9 years with Thales-Safran in the field of infra-red detectors development and manufacturing, as the technology reliability manager, and head of the Digital Factory. He spent also 5 years with IBM as a research engineer in the field of reliability of advanced Silicon Devices. He holds a master degree in Electonics and Telecommunications from *Ecole Supérieure de Chimie Physique Electronique de Lyon*, a Ph.D. in Microelectronics from the *University of Grenoble*, and a master degree in Data Science from *Ecole Polytechnique*, France. Augustin is an expert for the European Commission and member of the French standardization group for Artificial Intelligence.