A data-driven risk assessment approach for electronic boards used in oil well drilling operations

Delia-Elena Dumitru[1], Jinlong Kang[2], Alejandro Olid-Gonzalez[3] and Ahmed Mosallam[2]

[1] *SLB, Bucharest, 060201, Romania*

DDumitru2@slb.com

[2] *SLB, Clamart, 92140, France*

JKang5@slb.com

AMosallam@slb.com

[3] *SLB, Madrid, 28020, Spain*

AOlid@slb.com

ABSTRACT

To assist subject matter experts in investigating electronic failures of drilling tools, an innovative risk assessment approach for oil well drilling operations is developed that relies on synthetic time-series data to emulate environmental factors encountered downhole, explicitly focusing on temperature, shock, and vibration. The approach involves utilizing load cycle counting to extract meaningful features from each environmental channel measured by the drilling tool. The results from experiments with features related to dwell periods (dwell time and dwell damage) and load cycles (cycle means and cycle ranges) show a significant correlation between load cycle features and the risk label. Subsequently, a tree-based machine learning model is trained to label drilling operations based on synthetic data. Several models have been trained initially with comparable results. However, the advantage of using a tree-based model, specifically extra trees, is explainability and the stochastic aspect, which translates into model robustness when applied to real data. Preliminary results from a case study indicate that this new approach is highly effective in categorizing environmental risks associated with drilling operations. This risk assessment method can significantly enhance the decision-making process in investigating electronic board failures by offering reliable decision support.

1. INTRODUCTION

The drilling process has undergone a remarkable transformation in the oil and gas industry, evolving into a complex and sophisticated endeavor. This increased complexity stems directly from the necessity of accessing and extracting valuable resources hidden deep within the earth's crust. To accomplish this daunting task, the industry relies heavily on drilling tools, which are the technological cornerstones of these operations.

Drilling tools represent exceedingly intricate systems enriched with electronics, comprising a multitude of electronic boards, each meticulously designed to fulfill specialized functions of paramount significance to the success of drilling operations. These electronic boards function as the central hubs of technological operations, assuming responsibilities encompassing data acquisition, signal processing, management of control systems, and the facilitation of seamless communication (Kang et al., 2022). Thus, the reliability and performance of these electronic boards are inexorably linked to the overall effectiveness of drilling endeavors. However, the harsh operating conditions encountered downhole, including elevated temperatures, dynamic vibrations, and substantial shocks, render these boards susceptible to complex failure modes, potentially resulting in drilling operation failures. Failed drilling operations can lead to significant financial losses and environmental concerns. Therefore, health assessment and prognostics of electronic boards in drilling tools is essential to ensure that proactive maintenance is carried out in advance to

prevent drilling operations from failing.

The current health assessment and prognostics models for electronics are predominantly data-driven. For instance, physics-of-failure-based prognostics combine sensor data with models that evaluate a component's deviation from normal operation (Pecht & Gu, 2009). Another example is the use of accelerometers to monitor the response of printed circuit boards to vibrations and predict their remaining life (Gu, Barker, & Pecht, 2009). Similarly, (Vichare & Pecht, 2009) propose a technique that extracts load parameters from time-series data to estimate remaining life and assess damage. This method focuses on identifying valuable features for prognostics without requiring the storage of large volumes of data. Additionally, (Prisacaru, Gromala, Han, & Zhang, 2022) detect faults in electronic packages through the Mahalanobis distance and clarify them using a clustering technique. They also employ Echo state networks to perform degradation assessment and remaining useful life prediction. Additional literature on data-driven approaches for electronics health assessment and prognostics can be found in the following review articles: (Bhat, Muench, & Roellig, 2023), (Bhargava et al., 2020), and (Michael G. Pecht, Myeongsu Kang, 2018).

In the context of electronic boards used for oil well drilling operations, (Kale, Carter-Journet, Falgout, Heuermann-Kuehn, & Zurcher, 2014) propose a probabilistic approach that uses operational data, drilling dynamics, and historical maintenance information to predict reliability and life of electronics. (Bhatnagar, Cassou, Masry, & Mosallam, 2021) develop a data-driven fault detection approach tailored to electronic boards in intelligent remote dual-valve systems . Similarly, (V. Gupta et al., 2023) present an automatic fault detection method based on support vector machines for resistivity subsystems in Logging-While-Drilling (LWD) tools. (Sobczak-Oramus, Mosallam, Basci, & Kang, 2022) introduce a data-driven fault detection approach for transmitter subsystems in LWD tools . Finally, (Mosallam, Kang, Youssef, Laval, & Fulton, 2023) propose a data-driven fault diagnostics approach for three power supply boards in LWD tools.

Obtaining comprehensive data and corresponding labels throughout the equipment lifecycle is essential for building data-driven models for health assessment and prognostics of electronics. Subject matter experts usually derive data labels through failure investigations, but this process can be costly and time-consuming for complex equipment. Specifically, investigating electronic board failures in drilling tools requires manually examining extensive operational environment data measured by the tools. This process is labor intensive and prone to human error, making it challenging. Considering this challenge, this paper proposes an innovative risk assessment approach for oil well drilling operations to assist subject matter experts in investigating electronic failures. One of the primary advantages of this approach is its ability to harness the power of supervised learning for efficient and objective risk assessment, compared to manual inspections of operational environment data.

Literature has shown that various factors, such as temperature, humidity, vibration, dust, electrical stress, etc., affect the performance and life of electronic components (Michael G. Pecht, Myeongsu Kang, 2018). Among these factors, failures attributed to environmental conditions like temperature, humidity, and vibration constitute a significant 84% of electronic failures (Bhargava et al., 2020). Given the paramount importance of environmental factors in electronic failures, this paper seeks to develop a method to aid the subject matter experts investigate the specific environmental factors contributing to electronic failures.

However, only temperature and vibration are considered in the proposed method. We did not account for potential factors such as dust, humidity, chemicals, and radiation. This omission is because drilling tools do not typically measure these parameters for electronic boards. The physical arrangement of electronic boards within these tools inherently protects against exposure to dust, humidity, radiation, and chemicals that may be present in the wellbore. These tools are typically enclosed within robust steel tubing, shielding internal electronics from direct contact with these environmental factors. Moreover, before tool deployment, field engineers frequently introduce nitrogen into these tools, reducing the likelihood of exposure to potentially harmful substances. As a result of these protective measures and practices, the risk of electronic board damage due to dust, humidity, radiation, and chemical exposure is significantly mitigated.

The rest of this paper is structured into four sections. The first section offers a detailed presentation of the proposed method. Following that, a case study is presented. Finally, the last section summarizes the findings and suggests potential avenues for future research.

2. Proposed Method

The proposed method consists of three steps: data generation, preprocessing and feature extraction, and modelling, as illustrated in Figure 1.

2.1. Data generation

To leverage the power of supervised learning, labeled environmental data are needed. We generate synthetic time series programmatically to remove the need for expert-labeled data. Drilling tools regularly record measurements concerning the environment, specifically, temperature, shock peak values, and vibration root mean square values; therefore, in our experiment, we generate synthetic time series data that emulate drilling conditions for each of the three channels. The simulated data incorporate various sources of random-

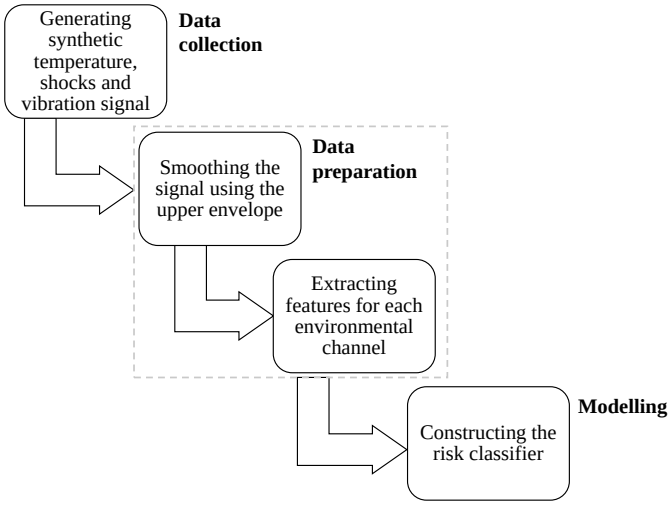


Figure 1. Proposed method.

ness, including sinusoidal waves with random time-variant amplitude and frequency, Gaussian noise, random spikes, and random shocks with random decay rates. Specifically, low-risk time series data exhibit lower parameter values for random number generation than high-risk time series data. For instance, the mean and standard deviation for generating low-risk temperature data's Gaussian noise are set to 40 and 3, respectively, while the amplitude for temperature shocks falls between 30 and 70. On the other hand, the mean and standard deviation for generating high-risk temperature data's Gaussian noise are set to 80 and 10, respectively, while the amplitude for temperature shocks falls between 50 and 100.

2.2. Data preparation

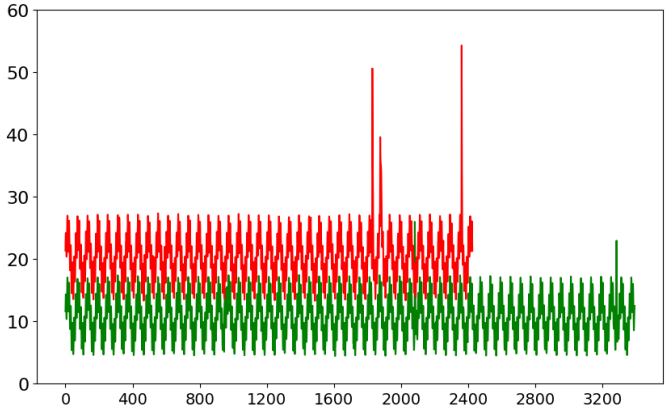
To effectively use the generated environmental data, preprocessing and optimal feature extraction are required. The preprocessing step consists of smoothing the signal using the upper envelope of the signal, as shown in Figure 2. After the preprocessing step, the environmental features can be extracted. For each environmental channel (i.e., temperature, shocks, and vibration) we compute two features based on dwell periods and two features based on load cycles, using the rainflow cycle counting method for the latter (Lee & Tjhung, 2012).

2.3. Feature extraction using rainflow cycle counting

Rainflow cycle counting is a method used in fatigue analysis to quantify the number of stress cycles experienced by a component or material (Endo, 1974).

The process involves analyzing a time series of stress or strain data to identify and count individual cycles. These cycles represent the repeated loading and unloading of a material, which can lead to fatigue failure over time. Rainflow cycle counting is especially useful for irregular or variable ampli-

(a) Raw signal



(b) Preprocessed signal

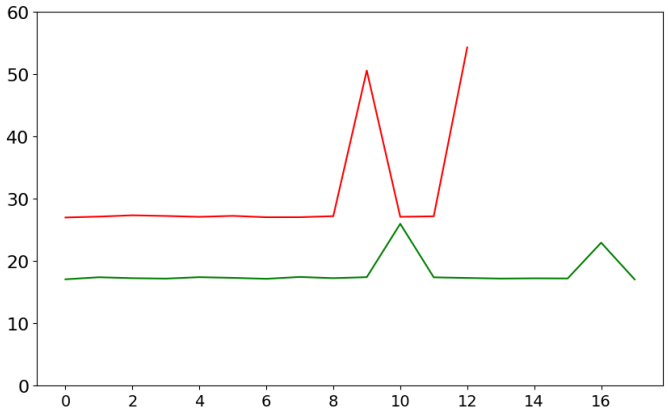


Figure 2. Generated vibration signal for a high-risk run (red) and a low-risk run(green), before and after preprocessing.

tude loading conditions, where the stress levels vary over time (Lee & Tjhung, 2012).

The method consists of four steps, as illustrated in Figure 3:

1. *Hysteresis filtering* (Figure 3a) entails removing cycles smaller than an amplitude gate that contribute minimal damage. This is accomplished by setting a gate with a specific amplitude. Any cycle with an amplitude below this gate is excluded from the load-time data. The gate is projected sequentially from left to right starting from each turning point in the time series. If a turning point falls below the gate's threshold, it is omitted from the time history. (Endo, 1974)(Lee & Tjhung, 2012).
2. *Peak-valley identification* (Figure 3b) consists of locating the points in the data where the direction of the signal reverses. In a cycle, only the highest and lowest values are pertinent for fatigue life assessments. Therefore, any intermediate data points between these extremes within a cycle can be disregarded as they do not contribute to the fatigue calculation. (Endo, 1974)(Lee & Tjhung, 2012).
3. In *discretization* (Figure 3c), the amplitude dimension of the signal is divided into a set number of equal bins. Each data point is then mapped to the center of its corresponding bin to facilitate cycle counting. Centering the data samples within their bins slightly modifies their amplitudes, therefore it is crucial to utilize an adequate number of bins in the analysis to minimize significant alterations in amplitudes (Endo, 1974)(Lee & Tjhung, 2012).
4. In *four-point counting* (Figure 3d), the identified peaks and valleys are connected to form hysteresis loops, or closed paths that represent complete stress cycles (Endo, 1974)(Lee & Tjhung, 2012). This is done using the following steps:
   (a) Select four consecutive points: $S1, S2, S3, S4$.
   (b) Compute inner stress: $|S2 - S3|$.
   (c) Compute outer stress: $|S1 - S4|$.
   (d) If the inner stress range is less than or equal to the outer stress range, a cycle is counted, otherwise it is not counted.

Using the method described above, the extracted features are as follows:

1. *average cycle mean*, where the cycle mean represents the mean values of the initial and final points of a cycle
2. *average cycle range*, where the cycle range represents the absolute difference between the initial and final points of a cycle
3. *dwell time*, representing the cumulative time during which the signal oscillation is lower than a set threshold
4. *dwell damage*, representing the average amplitude during the dwell time

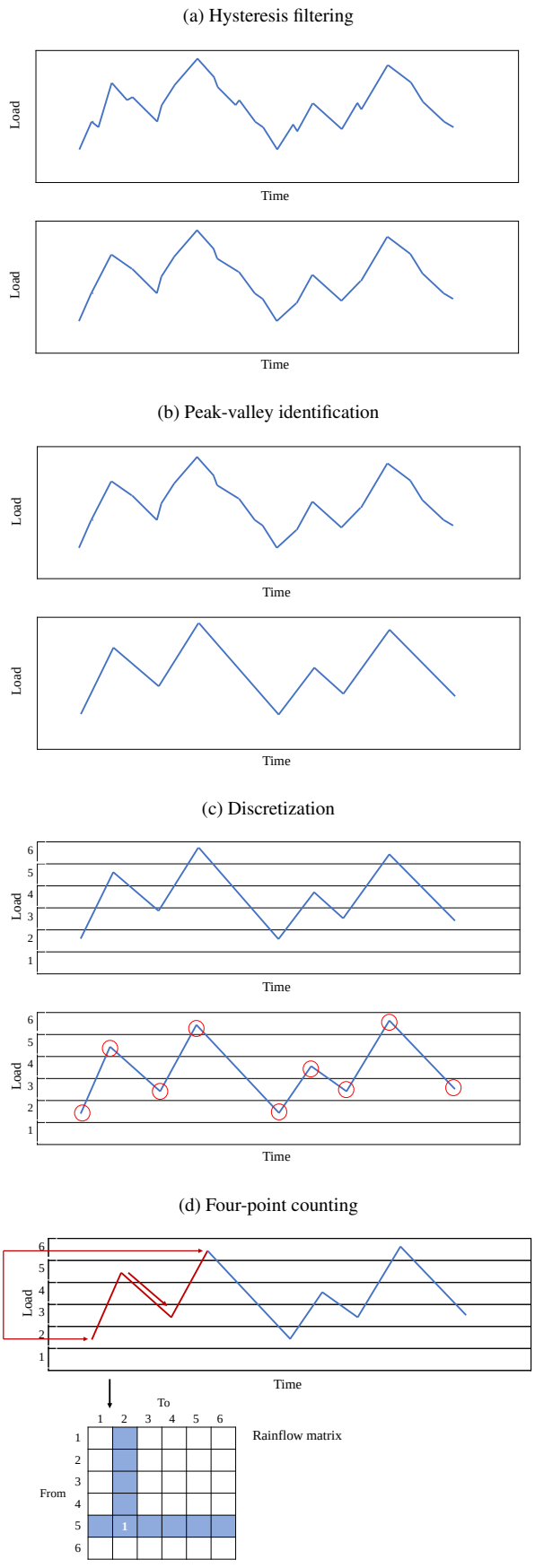


Figure 3. Rainflow cycle counting steps.

		Predicted label	
		Positive	Negative
Actual label	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

Figure 4. Confusion matrix for a binary classification problem.

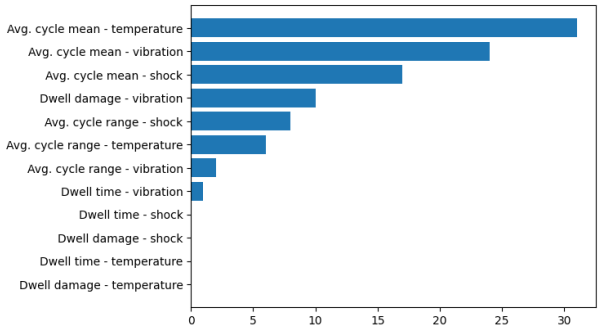


Figure 5. Model feature importance.

2.4. Modelling

We model the problem as a binary classification problem, where we interpret the positive class as high environmental risk, and the negative class as low environmental risk.

For risk classification three models were trained: logistic regression (LaValley, 2008), random forest (Biau & Scornet, 2016), and extra trees (Geurts, Ernst, & Wehenkel, 2006). The random forest and the extra trees models consist of an ensemble of 100 trees, and the Gini index was used as the splitting criterion. Logistic regression, as well as the ensemble tree-based models are less prone to overfitting and thus have the potential to generalize better to real data.

3. Case Study

A number of 1128 examples were generated, out of which 80% were used for training and 20% for testing. The training set was further split into train and validation sets in the same ratio using k-fold cross validation with 10 folds. The data were split as to preserve the class balance.

To evaluate the models on a labeled subset of the data we make use use of the confusion matrix (Fawcett, 2006), illustrated in Figure 4. In a binary classification problem, the confusion matrix has four sections:

1. True positives (TP): the number of instances where the model correctly predicts the positive class (high risk).
2. True negatives (TN): the number of instances where the model correctly predicts the negative class (low risk).
3. False positives (FP): the number of instances where the model incorrectly predicts the positive class.
4. False negatives (FN): the number of instances where the model incorrectly predicts the negative class.

To compare the models, we use the area under the receiver operating characteristic (ROC) curve (ROC AUC score). The ROC curve plots the true positive (TP) rate, defined as

$$\frac{TP}{TP+FN},$$

against the true negative (TN) rate, defined as

$$\frac{TN}{TN+FP}$$

It is a graphical representation of a binary classifier at different classification thresholds. The ROC AUC score is represented by the area under the ROC curve, where a score of 0.5 indicates a random model (Bradley, 1997).

The three trained models output a ROC AUC score of 1 on the validation set, indicated in Table 1. The application of the trained models is to assess environmental risk on electronic boards. Therefore, an important aspect is the ability of the model to successfully transfer knowledge from synthetic data to real data. In this regard, the stochastic features of the extra trees represent an advantage for increasing robustness (Geurts et al., 2006).

Table 1. Comparative ROC AUC score for the three trained models.

Model	ROC AUC score
Logistic regression	1.00
Random forest	1.00
Extra trees	1.00

We evaluate feature importance for the classification problem using Shapley values. This step helps to reduce feature redundancy and improve model interpretability. Shapley values are a method derived from cooperative game theory that has been adapted for use in explaining the predictions of machine learning models. They provide a way to fairly assess the impact of each feature for a particular prediction in a model (Merrick & Taly, 2020).

Using this method, Figure 5 indicates that for the extra trees model, the most impactful features are the average cycle means on each environmental channel, which is consistent with the feature correlation matrix in Figure 6.

Feature correlation in a machine learning model refers to the

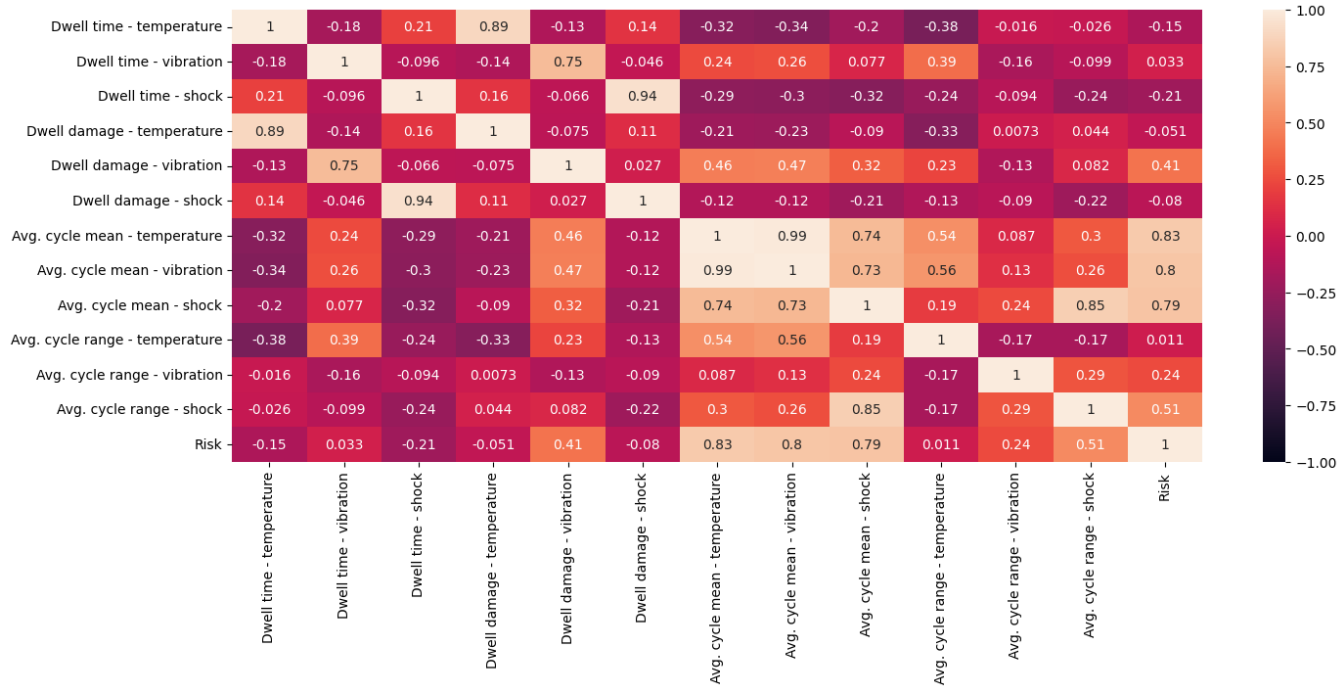


Figure 6. Feature correlation matrix.

degree to which the variables (features) in the dataset are related to each other, as well as with the target variable. For this experiment we use the Pearson correlation coefficient (Kendall & Stuart, 1973) and we specifically study the correlation between the features and the target variable, denominated as risk. In Figure 6 we notice the highest correlation between the average cycle means on the temperature, shock and vibration channels, and the risk variable.

In the second iteration of experiments, we restrict the training to these three features.

During the validation step, the model achieves promising classification results, as indicated by the confusion matrix in Table 2. Based on the confusion matrix, we define the following metrics:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

$$Precision = \frac{TP}{TP + FP}, \tag{2}$$

$$Recall = \frac{TP}{TP + FN}, \tag{3}$$

$$F1score = \frac{2 * TP}{2 * TP + FP + FN}. \tag{4}$$

The results on the validation set are consistent with the results on the test set after the training is completed, which can be seen in Table 3, despite the 0.52 score for data drift. Data drift indicates a difference in the statistical properties of the data. Therefore, the classification scores prove the robustness of the extra trees model and the potential for such a model to be used for assessing risk on real data.

Table 2. Confusion matrix on the validation set, where the positive class is equivalent to a high-risk run and the negative class is equivalent to a low-risk run.

	Predicted positive	Predicted negative
True positive	93	0
True negative	0	86

Table 3. Metrics measured on the test set.

Metric	Value
Accuracy	1.00
Precision	1.00
Recall	1.00
F1 Score	1.00
ROC AUC	1.00
Data drift	0.52

4. Conclusion and Future Work

This paper presented a data-driven approach for assessing environmental risk in electronic boards based on supervised machine learning. The method makes use of synthetic data and consists of extracting features with respect to dwell time and load cycles, showing that the latter have a larger impact on the performance of the models. The extra trees model achieves promising results on the synthetic data, but further work is needed to address the potential mismatch between training and test data in practical applications.

To address this issue, we plan to collect real-world environmental data and use it to fine-tune the model to better handle the variability of different environments. Additionally, we could explore the use of transfer learning techniques to adapt the model to new environments and improve its robustness to different types of data.

Overall, the proposed approach shows potential for assessing environmental risk in electronic boards, but further research is needed to optimize the model for real-world applications.

References

Bhargava, C., Sharma, P. K., Senthilkumar, M., Padmanaban, S., Ramachandaramurthy, V. K., Leonowicz, Z., ... Mitolo, M. (2020). Review of health prognostics and condition monitoring of electronic components. *IEEE Access*, *8*, 75163-75183. doi: 10.1109/ACCESS.2020.2989410

Bhat, D., Muench, S., & Roellig, M. (2023). Application of machine learning algorithms in prognostics and health monitoring of electronic systems: A review. *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, *4*, 100166. doi: 10.1016/j.prime.2023.100166

Bhatnagar, S., Cassou, M. L., Masry, Z. A., & Mosallam, A. (2021, June). Data-Driven Fault Detection Method for Electronic Boards in Intelligent Remote Dual-Valve System. In *PHM Society European Conference* (pp. 1–7). doi: 10.36001/phme.2021.v6i1.2903

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*, 197–227.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145-1159. doi: 10.1016/S0031-3203(96)00142-2

Endo, T. (1974). Damage evaluation of metals for random or varying loading. In *Proceedings of the 1974 Symposium on Mechanical Behavior of Materials* (p. 371-380).

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, *27*(8), 861-874. Retrieved from https://www.sciencedirect.com/science/article/pii/S016786550500303X (ROC Analysis in Pattern Recognition) doi: https://doi.org/10.1016/j.patrec.2005.10.010

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, *63*, 3–42.

Gu, J., Barker, D., & Pecht, M. (2009). Health monitoring and prognostics of electronics subject to vibration load conditions. *IEEE Sensors Journal*, *9*(11), 1479-1485.

Kale, A., Carter-Journet, K., Falgout, T., Heuermann-Kuehn, L., & Zurcher, D. (2014). A probabilistic approach for reliability and life prediction of electronics in drilling and evaluation tools. In *Proceedings of the Annual Conference of the Prognostics and Health Management Society 2014* (p. 519-532).

Kang, J., Varnier, C., Mosallam, A., Zerhouni, N., Youssef, F. B., & Shen, N. (2022). Risk level estimation for electronics boards in drilling and measurement tools based on the hidden Markov model. In *2022 Prognostics and Health Management Conference (PHM-2022 London)* (p. 495-500). doi: 10.1109/PHM2022-London52454.2022.00093

Kendall, M., & Stuart, A. (1973). *The advanced theory of statistics. vol. 2: Inference and: Relationsship*. Griffin.

LaValley, M. P. (2008). Logistic regression. *Circulation*, *117*(18), 2395–2399.

Lee, Y.-L., & Tjhung, T. (2012). Chapter 3 - rainflow cycle counting techniques. In Y.-L. Lee, M. E. Barkey, & H.-T. Kang (Eds.), *Metal fatigue analysis handbook* (p. 89-114). Boston: Butterworth-Heinemann. doi: 10.1016/B978-0-12-385204-5.00003-3

Merrick, L., & Taly, A. (2020). The explanation game: Explaining machine learning models using shapley values. In A. Holzinger, P. Kieseberg, A. M. Tjoa, & E. Weippl (Eds.), *Machine learning and knowledge extraction* (pp. 17–38). Cham: Springer International Publishing.

Michael G. Pecht, Myeongsu Kang. (2018). *Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things*. John Wiley and Sons Ltd.

Mosallam, A., Kang, J., Youssef, F. B., Laval, L., & Fulton, J. (2023, May). Data-Driven Fault Diagnostics for Neutron Generator Systems in Multifunction Logging-While-Drilling Service. In *2023 Prognostics and Health Management Conference (PHM)* (pp. 171–176). doi: 10.1109/PHM58589.2023.00041

Pecht, M., & Gu, J. (2009). Physics-of-failure-based prognostics for electronic products. *Transactions of the Institute of Measurement and Control*, *31*(3-4), 309-322. doi: 10.1177/0142331208092031

Prisacaru, A., Gromala, P., Han, B., & Zhang, G. Q. (2022). Degradation estimation and prediction of electronic packages using data-driven approach. *IEEE Transactions on Industrial Electronics*, *69*(3), 2996-3006. doi:

10.1109/TIE.2021.3068681

Sobczak-Oramus, K., Mosallam, A., Basci, C., & Kang, J. (2022, June). Data-Driven Fault Detection for Transmitter in Logging-While-Drilling Tool. In *PHM Society European Conference* (Vol. 7, pp. 458–465). doi: 10.36001/phme.2022.v7i1.3362

V. Gupta, J. Kang, A. Mosallam, N. Shen, F. B. Youssef, & L. Laval. (2023, June). Automatic Fault Detection for Resistivity Systems in Logging-While-Drilling Tools. In *2023 Prognostics and Health Management Conference (PHM)* (pp. 128–132). doi: 10.1109/PHM58589.2023.00032

Vichare, N., & Pecht, M. (2009). *Method to extract parameters from in-situ monitored signals for prognostics* (No. US8521443B2).

BIOGRAPHIES

Delia-Elena Dumitru Delia-Elena Dumitru is a Data Scientist at the SLB IT center in Bucharest, Romania. She graduated in 2018 with a B.S. degree in Computer Science and completed her M.S. degree in Applied Computational Intelligence in 2020, both at the Babeș-Bolyai University in Cluj-Napoca, Romania. Her main research interests are machine learning, computer vision and PHM.

Jinlong Kang is currently a data scientist at SLB technology center in Clamart, France. He received the B.S. degree in Industrial Engineering in 2016 and the M.S. degree in Mechanical Engineering in 2019 both from University of Electronic Science and Technology of China, and the Ph.D. degree in automatics in 2024 from University of Franche-Comté in France. His main research interests are Prognostic and Health Management (PHM), maintenance decision-making, data mining and machine learning.

Ahmed Mosallam is the Data Science AI European Hub Manager at SLB technology center in Clamart, France. He has his Ph.D. degree in automatic control in the field of PHM from University of Franche-Comté in Besançon, France. His main research interests are signal processing, data mining, machine learning and PHM.

Alejandro Olid Gonzalez is currently a data scientist at SLB in Madrid, Spain. He obtained his B.Sc. degree in Physics in 2013 and his M.Sc. degree in Astrophysics in 2017. His current research interests are Prognostic and Health Management (PHM), machine learning, and simulations of physical systems.