

Machine Learning for Combined Fault Diagnosis in Rotating Machines From Training With Simple Faults

Flavio Henrique Origuela Meira¹, Dionísio Henrique Carvalho de Sá Só Martins², and Diego Barreto Haddad¹

Abstract—This paper presents a machine learning framework for diagnosing simple and combined mechanical faults in rotating machines from vibration data. The main goal is to investigate whether a classifier trained mostly on simple faults can generalize to previously unseen fault combinations. The study uses the COMFAULDA dataset, whose instances are represented by 217 features extracted from the time and frequency domains. Four classifiers are evaluated under a binary multi-target formulation: multilayer perceptron, Random Forest, Extreme Gradient Boosting, and support vector machine. Five experimental scenarios are defined, from zero exposure to combined faults during training to progressive partial exposure. Results show that pure compositional generalization is difficult, but controlled inclusion of a small fraction of combined-fault samples significantly improves performance. The best result was obtained by XGB with Bayesian hyperparameter optimization, achieving a combined F_1 score of 0.4489 and an overall accuracy of 0.8796 in the most favorable scenario. The results indicate that the proposed framework is promising for predictive maintenance applications in which combined faults are rare or only partially labeled.

Keywords—Fault diagnosis, rotating machines, multi-target classification, predictive maintenance.

I. INTRODUCTION

Fault diagnosis in rotating machines is essential for ensuring reliability, safety, and cost-effective maintenance in industrial systems. Vibration-based condition monitoring has become a widely adopted approach due to its noninvasive nature and ability to detect early-stage faults [1]. With the advent of Industry 4.0, machine learning (ML) techniques have increasingly been applied to enhance diagnostic performance through data-driven analysis.

Most existing studies focus on the classification of *simple faults*, where each condition corresponds to a single defect. Under this assumption, supervised models such as support vector machine, neural networks, and ensemble methods achieve high accuracy. However, these approaches rely on the independence of faults, which rarely holds in practice [2]. In real environments, multiple faults may occur simultaneously, producing overlapping or interacting signatures.

Combined faults introduce additional complexity, as their vibration patterns differ from those of isolated defects and

are often underrepresented in datasets. The scarcity of labeled data for such conditions limits the generalization capability of models trained only on simple faults [2]. This motivates the investigation of strategies that allow models to infer fault combinations without explicit exposure during training.

In recent years, several machine learning algorithms have been successfully applied to vibration-based fault diagnosis in rotating machinery [1], [2]. The support vector machine (SVM) is widely adopted due to its strong generalization capability and effectiveness in high-dimensional feature spaces, whereas the Random Forest (RF) has demonstrated robust performance by aggregating multiple decision trees [2], [3].

More recently, machine learning approaches, including Extreme Gradient Boosting (XGB) and the multilayer perceptron (MLP), have been applied to fault diagnosis in rotating machinery. These models are capable of capturing nonlinear relationships in vibration signals and, when combined with feature extraction techniques in the time and frequency domains, provide effective solutions for predictive maintenance [3], [4].

In this context, this paper investigates whether models trained exclusively on simple faults can generalize to *combined fault diagnosis* using a multi-target formulation. Experiments are conducted on the COMFAULDA dataset [5], which provides a controlled experimental environment suitable for evaluating diagnostic generalization under different levels of combined-fault supervision. Different classifiers are evaluated under varying levels of exposure to combined faults. The study aims to assess the feasibility and limitations of compositional generalization for predictive maintenance applications [6].

The remainder of this paper is organized as follows. Section II presents the methodology, Section III discusses the experimental results, and Section IV concludes the paper.

II. METHODOLOGY

The experiments are conducted on the COMFAULDA dataset [5], which contains 2,162 recorded vibration signals acquired from a controlled rotating-machine test bench and distributed among seven operating conditions.

Figure 1 illustrates the experimental setup used for data acquisition, including the motor, shaft, and sensor positioning.

The dataset includes seven operating conditions: normal operation, unbalance, horizontal misalignment, vertical misalignment, unbalance combined with horizontal misalignment, unbalance combined with vertical misalignment, and simulta-

¹Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ), Brazil, ² Instituto Nacional da Propriedade Industrial (INPI). E-mails: flavio.meira@cefet-rj.br, diego.haddad@cefet-rj.br, dionisio.martins@inpi.gov.br. This paper was supported in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), in part by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior Brasil (CAPES, Finance Code 001), and in part by Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ).

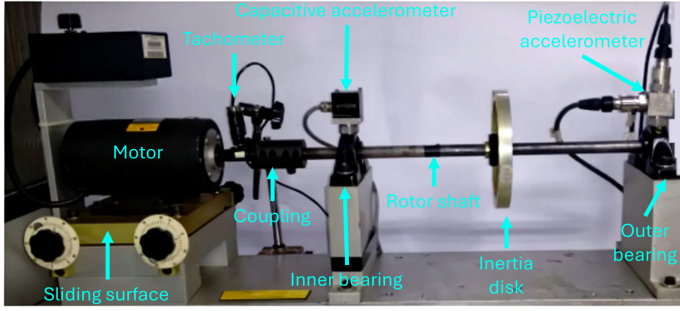


Fig. 1. Experimental test bench used for vibration data acquisition in the COMFAULDA dataset (adapted from [5]).

neous horizontal and vertical misalignment. These fault scenarios were artificially induced, enabling a systematic evaluation of isolated and combined fault behaviors.

Each vibration record is represented by 217 features extracted from signals measured along the X, Y, and Z directions. The time-domain features include maximum and minimum values, mean, variance, standard deviation, root mean square, kurtosis, skewness, signal energy, amplitude range, entropy, crest factor, shape factor, impulse factor, margin factor, and defect factor. The frequency-domain features include analogous statistical descriptors calculated from the spectrum, spectral energy, dominant spectral components, and amplitudes associated with the shaft rotational frequency and its harmonics. Additional descriptors include energy and local peaks computed over spectral sub-bands, peak-to-noise relationships, spectral-variation measures, and rotation-related information obtained from the tachometer. The features extracted from the three vibration directions are concatenated to form the final input vector used by the classifiers [1], [5].

Our methodology is designed to systematically investigate the ability of machine learning models to perform compositional generalization, i.e., to infer previously unseen fault combinations from training data containing predominantly simple faults. The framework encompasses model selection, multi-target formulation, training protocol, controlled exposure scenarios, and evaluation metrics.

A. Learning Models and Hyperparameter Optimization

Four supervised learning algorithms were evaluated: MLP, RF, XGB, and SVM. These models were selected to provide complementary inductive biases and represent distinct families of learning paradigms.

Tree-based ensemble methods such as RF and XGB are capable of capturing nonlinear feature interactions through hierarchical partitioning of the input space [2], [3]. In contrast, MLP and SVM provide nonlinear classification capabilities that can be applied to complex fault interactions in vibration signals [2], [4].

The multi-fault problem is formulated as independent binary classification tasks, one for each elementary fault component. Given $\mathbf{x} \in \mathbb{R}^d$, the model predicts $\mathbf{t} \in \{0, 1\}^K$, where each dimension indicates the presence of a fault. This multi-target representation expresses combined conditions as compositions

of elementary faults and supports the inclusion of new combinations.

The four-bit target vector follows the order [vertical misalignment, horizontal misalignment, unbalance, normal operation]. Accordingly, the binary labels used in the confusion matrices are defined as follows: 0001 denotes normal operation, 0010 denotes unbalance, 0100 denotes horizontal misalignment, 1000 denotes vertical misalignment, 0110 denotes unbalance combined with horizontal misalignment, 1010 denotes unbalance combined with vertical misalignment, and 1100 denotes simultaneous horizontal and vertical misalignment.

Hyperparameter optimization was performed using three search strategies: grid search, random search, and Bayesian optimization. These approaches were selected to balance systematic exploration and computational efficiency. Grid search evaluates predefined combinations of hyperparameters, whereas random search explores configurations through random sampling and can be more efficient when only a subset of hyperparameters has a strong influence on model performance [7]. Bayesian optimization uses a probabilistic surrogate model and an acquisition function to guide the search toward promising regions of the hyperparameter space [8].

The optimized hyperparameters of the final best-performing configuration, corresponding to XGB with Bayesian optimization in Scenario C4, are presented in Table I. Hyperparameter selection was independently performed for each binary target using the F1 score as the cross-validation criterion.

TABLE I
OPTIMIZED HYPERPARAMETERS OF THE FINAL XGB CONFIGURATION IN SCENARIO C4.

Target	N	D	η	S	C	F1-CV
F1	499	3	0.0487	0.7564	0.9487	0.9302
F2	227	37	0.3801	0.9210	0.8826	0.8477
F3	226	47	0.1723	0.8351	0.7051	0.9505
F4	332	48	0.0634	0.6513	0.5046	0.9512

N : number of estimators; D : maximum depth; η : learning rate; S : subsample ratio; C : column-sampling ratio.

Model selection prioritized combined-fault performance over overall accuracy, consistently with the objective of the study.

B. Data Splitting and Training Protocol

The dataset was divided into training, validation, and test sets using stratified hold-out to preserve fault-type distributions.

In addition, 3-fold cross-validation is employed within the training set during hyperparameter tuning to ensure robust performance estimation and reduce variance associated with data partitioning. All preprocessing steps, including feature standardization, are performed using statistics computed exclusively from the training data to avoid data leakage.

Simple-fault samples were always available for training and validation, whereas combined-fault samples were progressively introduced according to the following scenarios.

C. Controlled Exposure Scenarios

To systematically assess the impact of supervision on combined fault recognition, five experimental scenarios were defined, denoted as C0 through C4. Each scenario specifies the proportion of combined-fault samples allocated to training, validation, and test sets.

TABLE II
DISTRIBUTION OF COMBINED-FAULT SAMPLES ACROSS EXPERIMENTAL SCENARIOS.

Scenario	Training	Validation	Test
C0	0%	0%	100%
C1	5%	1%	94%
C2	5%	5%	90%
C3	10%	10%	80%
C4	20%	10%	70%

The allocation of simple-fault samples was kept fixed across all experimental scenarios, whereas only the distribution of combined-fault samples was modified. Therefore, the total number of test samples decreases progressively from C0 to C4 as combined-fault instances are transferred from the test set to the training and validation sets. No additional simple-fault samples were added to compensate for this reduction. Consequently, comparisons across scenarios are based on the same fault categories and evaluation metrics, but not on test sets with identical sample sizes.

This design was intentionally adopted to measure the effect of increasing exposure to combined faults while preserving a substantial independent combined-fault test subset in every scenario. Nevertheless, the variation in test-set size constitutes a limitation of the experimental protocol and should be considered when comparing absolute results across scenarios.

The five scenarios were selected as representative and interpretable levels of exposure to combined faults, rather than as an exhaustive evaluation of all possible proportions. Scenario C0 establishes the complete no-exposure baseline, whereas C1 and C2 evaluate the effect of introducing a minimal amount of combined-fault information. Scenario C3 represents an intermediate level of supervision, and C4 represents the highest exposure considered while still preserving most combined-fault samples for independent testing. The adopted proportions were chosen to analyze the transition from diagnosis without combined-fault examples to progressively supervised recognition without excessively reducing the combined-fault test set.

Scenario C0 represents a zero-exposure condition, in which the model has no prior exposure to combined faults. In this case, successful predictions rely entirely on the ability to recombine knowledge learned from simple faults.

Scenarios C1 to C4 progressively introduce limited amounts of combined-fault samples during training and validation, characterizing progressively supervised exposure to combined-fault conditions. This gradual increase enables the analysis of how small amounts of supervised data influence the model's ability to capture interactions between fault components.

The scenarios represent three regimes: pure compositional generalization (C0), low-supervision adaptation (C1–C2), and increased supervision with potential saturation (C3–C4).

D. Evaluation Metrics

The primary metric, $F1_{\text{comb}}$, balances precision and recall considering only samples containing two or more simultaneous faults.

Overall classification accuracy and the global macro-averaged F1 score are reported as complementary metrics to assess performance across all operating conditions. Accuracy measures the proportion of correctly classified instances, whereas macro-F1 assigns equal importance to each operating condition by averaging the class-wise F1 scores. The emphasis on $F1_{\text{comb}}$ reflects the primary objective of this work, while the global metrics provide a broader assessment of model performance.

The multi-target formulation also enables partial correctness, since individual fault components may be identified even when the complete combination is misclassified.

E. Methodological Considerations

The proposed methodology is aligned with realistic industrial constraints, where labeled data for combined faults is scarce or incomplete. By explicitly controlling the level of exposure to such conditions, the framework enables a detailed analysis of the trade-off between data availability and diagnostic performance.

It should be noted that the simple and combined faults in the COMFAULDA dataset were artificially induced under controlled laboratory conditions. This procedure enables repeatable experiments and systematic control of fault types and severities; however, the resulting vibration signatures may not fully reproduce the coupled degradation mechanisms, operational variability, load fluctuations, environmental disturbances, and progressive wear observed in real industrial machinery. Therefore, the results should be interpreted as evidence obtained under controlled experimental conditions rather than as a complete representation of naturally evolving industrial faults.

Validation using field data collected from industrial rotating machines is therefore required before the proposed framework can be generalized to real predictive-maintenance applications.

Furthermore, the use of a multi-target representation promotes interpretability and modularity, allowing the system to be extended to additional fault types without requiring a complete redefinition of the label space.

Overall, the methodology provides a structured and reproducible approach for evaluating compositional generalization in fault diagnosis systems, bridging the gap between idealized laboratory conditions and practical deployment scenarios.

III. RESULTS AND DISCUSSION

The experimental results assess the ability of the evaluated models to recognize combined faults under different levels of supervision. As summarized in Table II, performance strongly depends on the proportion of combined-fault samples available during training.

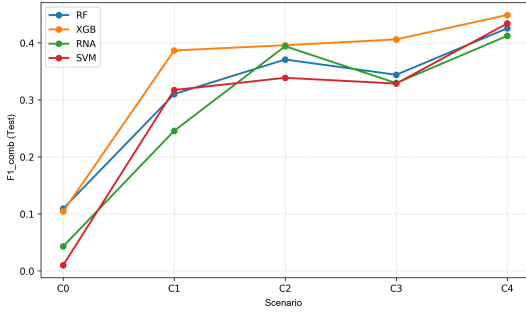


Fig. 2. Evolution of the F1 score for combined faults ($F1_{\text{comb}}$) across the experimental scenarios (C0–C4) for the evaluated models using Bayesian optimization.

A. Global Performance Analysis

Figure 2 compares model performance under progressive exposure to combined-fault samples.

In Scenario C0, where no combined-fault samples are available during training or validation, all models exhibit limited performance. The best configuration achieved $F1_{\text{comb}} = 0.1092$ and an overall accuracy of 0.2806, confirming that pure compositional generalization is highly challenging in this feature space.

This result indicates that, although the multi-target representation provides a theoretically suitable framework for compositional reasoning, the models are not able to fully reconstruct the joint distribution of fault interactions based solely on isolated patterns. In practice, the learned decision boundaries tend to approximate marginal fault behaviors, but fail to capture higher-order dependencies.

A significant improvement is observed in Scenario C1, where only 5% of combined-fault samples are introduced during training. In this case, $F1_{\text{comb}}$ increases to 0.3313, representing a relative improvement of more than 200% with respect to C0. This result highlights the importance of even minimal supervision in guiding the model toward meaningful representations of fault interactions.

From C1 to C4, the improvement trend is monotonic. In Scenario C4, XGB with Bayesian optimization achieves the best performance, reaching $F1_{\text{comb}} = 0.4489$, an accuracy of 0.8796, and a global macro-F1 score of 0.8515. This progression indicates that the inclusion of combined-fault samples enables the model to progressively refine its internal representation of nonlinear interactions between features.

It is also important to note that accuracy increases more rapidly than $F1_{\text{comb}}$, suggesting that simple faults remain easier to classify across all scenarios. Therefore, $F1_{\text{comb}}$ remains the most informative metric for assessing the core challenge addressed in this work.

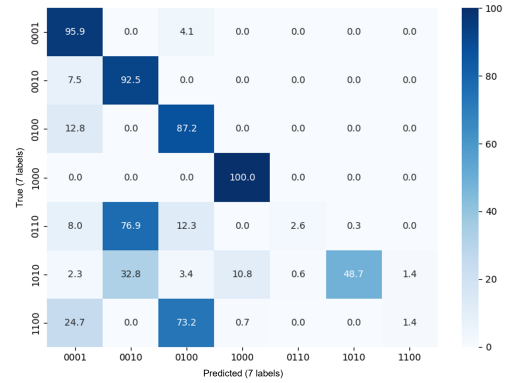
B. Impact of Controlled Exposure

The largest gain occurs from C0 to C1, indicating that the main limitation is the absence of examples encoding fault interactions. C0 requires extrapolation beyond the training distribution, whereas even limited exposure in C1 provides reference patterns that support interpolation.

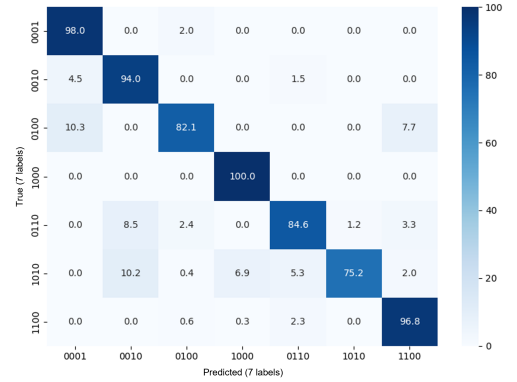
From C2 to C4, gains become more gradual, suggesting diminishing returns as the most relevant interactions are learned. More expressive models, particularly XGB, benefit more consistently from additional combined-fault samples. These results support a hybrid strategy combining abundant simple-fault data with a limited number of combined-fault examples.

C. Confusion Matrix Analysis

Fig. 3 compares the confusion matrices of the best-performing model (XGB) in C0 and C4.



(a) Scenario C0 (no exposure to combined faults)



(b) Scenario C4 (highest exposure)

Fig. 3. Confusion matrices of the XGB model with Bayesian optimization. The binary labels correspond to the seven operating conditions defined in Section II.

In C0, combined faults are frequently decomposed into individual components, indicating that the model recognizes elementary signatures but learns an approximately additive representation. In C4, supervised exposure substantially reduces these errors and improves the recognition of joint fault patterns.

Some confusion remains between similar combinations, particularly horizontal and vertical misalignment, due to overlapping feature representations. Nevertheless, the multi-target formulation often identifies individual components even when the complete combination is incorrect, providing useful partial diagnostic information.

D. Model Comparison

Among the evaluated models, XGB shows strong performance under partial exposure (C2–C4) and achieves the best overall result in Scenario C4. This can be attributed to its ability to model complex feature interactions through boosted decision trees and its robustness to heterogeneous feature distributions.

Random Forest also demonstrates stable performance, but its averaging mechanism limits its ability to capture fine-grained interactions compared to boosting methods.

SVM performs competitively in low-exposure scenarios, particularly in C1, likely due to its strong generalization properties and margin-based learning. However, its performance does not scale as effectively with increased data, possibly due to limitations in modeling complex nonlinear interactions in high-dimensional feature spaces.

The MLP exhibits higher variance and sensitivity to hyperparameter tuning, possibly due to the dataset size and the use of handcrafted features.

Overall, the results indicate that XGB provided the best predictive performance among the evaluated models under the adopted experimental protocol. However, this conclusion is based on classification metrics and does not account for computational cost. Training and inference times were not systematically benchmarked under a common hardware and software environment; therefore, the present results do not by themselves establish the suitability of XGB for real-time deployment.

E. Practical Implications

The main practical implications for predictive maintenance are:

- Pure compositional generalization is insufficient for reliable combined-fault diagnosis in practical scenarios.
- A small, carefully curated dataset of combined faults can significantly enhance model performance.
- Multi-target representations enable interpretable and partially correct predictions, which are valuable for diagnostic support.
- Tree-based ensemble models, particularly **XGB**, are well suited for capturing nonlinear fault interactions.

For practical deployment, predictive performance must be considered together with computational requirements. In particular, training time, model size, memory consumption, and inference latency should be measured on the intended processing platform. Since these aspects were not systematically compared in the present study, real-time applicability remains to be experimentally demonstrated.

The limited performance in Scenario C0 suggests that the handcrafted features may not fully represent the physical interactions between simultaneous faults.

In practical rotating machinery, combined faults are typically the result of interacting mechanical phenomena, such as imbalance and misalignment occurring simultaneously. These interactions generate coupled vibration signatures that are not simply additive, leading to complex patterns that are more difficult to distinguish than those of isolated faults. This

characteristic reinforces the importance of models capable of capturing nonlinear feature interactions, which is consistent with the superior performance observed for ensemble-based methods in this study [1], [2].

Despite these limitations, even limited exposure to combined faults substantially improves performance, which is relevant when labeled examples of all combinations are impractical. Hyperparameter optimization also supports model robustness and generalization [7], [8]. Nevertheless, generalization remains dependent on training-data diversity: although the multi-target formulation enables partial fault identification, the models struggle to reconstruct unseen combinations, consistent with compositional learning studies [6]. This challenge is intensified by nonlinear interactions between mechanical faults [1], [2].

IV. CONCLUSION

This paper investigated whether models trained mostly on simple faults can generalize to combined-fault diagnosis using the COMFAULDA dataset, 217 vibration features, and a binary multi-target representation.

Pure compositional generalization was limited, whereas limited exposure to combined-fault samples produced substantial gains. XGB with Bayesian hyperparameter optimization achieved the best result in C4, with $F1_{\text{comb}} = 0.4489$, accuracy of 0.8796, and global macro-F1 of 0.8515.

The limited performance in C0 suggests that handcrafted features may not fully represent nonlinear fault interactions, motivating learned or physics-informed representations.

Future work includes validation with industrial field data and additional operating conditions, investigation of feature selection and interpretability, and systematic benchmarking of training time, inference latency, model size, and memory consumption before embedded or real-time deployment.

REFERENCES

- [1] H. Yang, J. Mathew, and L. Ma, "Vibration feature extraction techniques for fault diagnosis of rotating machinery: a literature survey," in *Asia-pacific vibration conference*, no. 42460, pp. 801–807, 2003.
- [2] D. H. C. de Sá Só Martins, D. P. Viana, A. A. de Lima, M. F. Pinto, L. Tarrataca, F. Lopes e Silva, R. H. R. Gutiérrez, T. de Moura Prego, U. A. B. V. Monteiro, and D. B. Haddad, "Diagnostic and severity analysis of combined failures composed by imbalance and misalignment in rotating machines," *The International Journal of Advanced Manufacturing Technology*, vol. 114, no. 9, pp. 3077–3092, 2021.
- [3] J. Bogatinovski, L. Todorovski, S. Džeroski, and D. Kocev, "Comprehensive comparative study of multi-label classification methods," *arXiv preprint arXiv:2102.07113*, 2021.
- [4] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.
- [5] D. Martins, D. Pestana-Viana, A. Lima, D. Haddad, R. Homero, and L. Vaz, "Composed Fault Dataset (COMFAULDA)," *IEEE Dataport*, Nov. 2022.
- [6] J. Nam, J. Kim, E. Loza Mencía, I. Gurevych, and J. Fürnkranz, "Large-scale multi-label text classification—revisiting neural networks," in *Joint european conference on machine learning and knowledge discovery in databases*, pp. 437–452, Springer, 2014.
- [7] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *The journal of machine learning research*, vol. 13, no. 1, pp. 281–305, 2012.
- [8] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," *Advances in neural information processing systems*, vol. 25, 2012.