signals

Article

Machine Learning-Based Oil Analysis for Underground Mining Equipment

Nelson Chambi *, **Celso Sanga**, **Alejandra Sanga** and **Piero Sanga**

Facultad de Ingeniería de Producción y Servicios, Universidad Nacional de San Agustín de Arequipa, Arequipa 04000, Peru; csanga@unsa.edu.pe (C.S.); asanga@unsa.edu.pe (A.S.); psangah@unsa.edu.pe (P.S.)
* Correspondence: nchambiq@unsa.edu.pe; Tel.: +51-948502806

Abstract

Predictive maintenance in underground mining faces challenges due to severe conditions such as confined environments, high humidity, presence of silica dust, and restricted access. This study develops a predictive framework based on oil analysis and machine learning for multiple compartments of mining equipment (engine, hydraulic system, transmission, differential). Samples were processed under ASTM standards, integrating wear metal concentrations (Fe, Cu, Cr, Pb, Al), physicochemical properties (viscosity, TBN, soot), and contaminants (Si, Na). Based on tribology, interpretable ratios were constructed. Three algorithms (Random Forest, Gradient Boosting, and XGBoost) were evaluated using cross-validation. XGBoost achieved the best balance (F1 = 0.852, AUC = 0.975), with a recall of 94.5% for the critical class and only 3 false negatives out of 199 test samples, while Random Forest presented the highest global discrimination power (AUC = 0.978). SHAP revealed that viscosity at 100 °C is the most important predictor (SHAP ~0.9), surpassing iron. No temporal wear trend was found (R^2 = 0.000). Threshold optimization to 0.25 reduced false negatives by 67% (from 9 to 3). The framework provides interpretable predictions with uncertainty quantification for underground environments.

Keywords: predictive maintenance; oil analysis; machine learning; underground mining

1. Introduction

Operational continuity in underground mining critically depends on the performance of diesel engines installed in high-relevance equipment, such as scoops (LHDs), low-profile trucks, drilling jumbos, and haulage equipment. The failure of these engines not only causes unplanned shutdowns with significant production losses but also compromises personnel safety, process stability, and global operational availability. This problem intensifies in the underground mining context due to severe working conditions, characterized by confined environments, high humidity, suspended silica dust, variable operational loads, persistent mechanical vibration, and restricted access for inspection and repair tasks. Consequently, conventional corrective and preventive maintenance approaches become progressively less adequate. The former reacts once a failure has occurred, while the latter intervenes at fixed periodicities that do not reflect the actual degradation state of the asset.

In response, data-driven predictive maintenance has gained relevance as a strategy capable of anticipating failures through continuous monitoring of operational variables. Artificial intelligence has particularly expanded the capabilities of predictive maintenance by enabling the modeling of non-linear relationships, extracting complex patterns from temporal signals, and estimating degradation indicators difficult to capture using classical

Academic Editors: Chaoyong Ma, Kun Zhang, Yanping Zhu and Yonggang Xu

Received: 18 April 2026
Revised: 4 June 2026
Accepted: 11 June 2026
Published: 18 June 2026

techniques. This evolution has been driven by the increasing availability of sensors, SCADA platforms, IoT infrastructure, and advanced machine and deep learning techniques [1–4]. For example, in the manufacturing sector, models such as LightGBM have achieved 93% accuracy in early machine fault detection [5]. However, the effective application of these advances to underground mining equipment engines remains limited. The main barriers include data heterogeneity, noise contamination, scarcity of labeled records, variability in operating regimes, and the difficulty of transferring models developed in other sectors to real mining environments [6,7].

Moreover, although promising results have been reported in heavy machinery, bearings, generators, and even energy systems, a significant portion of the literature focuses on isolated components, controlled scenarios, or high-level reviews. In open-pit mining, XGBoost models optimized using particle swarm algorithms have predicted the failure time of mechanical shovels with a coefficient of determination of 0.99 [8]. This predictive capability has also been explored in other mining sectors, such as fatigue failure prediction in lifting systems using digital twins [9]. In the renewable energy sector, ensemble approaches combining Random Forest, XGBoost, and LightGBM have achieved 99.99% accuracy in detecting bearing failures in wind turbines using millions of high-frequency vibration samples [10]. For critical mining machinery, deep learning architectures such as CNN and LSTM have enabled real-time fault diagnosis and degradation assessment through the processing of vibration and electrical current signals [11]. Likewise, in industrial transportation systems, motor current signature analysis integrated with LSTM models has enabled precise prediction of mechanical failures in conveyor belts, facilitating prescriptive maintenance actions [12]. Recent initiatives in the autonomous vehicle sector for mining have also demonstrated the potential of variational autoencoders to detect anomalies in CAN data buses [13].

Nevertheless, a gap persists between algorithmic accuracy in controlled environments and the practical utility of models under real operational constraints. In the aerospace sector, hybrid architectures such as 1D-CNN-Informer with Monte Carlo dropout have demonstrated that probabilistic remaining useful life (RUL) predictions can be integrated with inventory optimization systems to reduce costs [14]. Along the same lines, approaches combining Restricted Boltzmann Machines (RBM) and Bi-LSTM networks have predicted the RUL of turbofan engines with an RMSE of 13.28 [15]. In heavy industry and manufacturing, integrated dynamic decision-making models have balanced inventory costs against stockout risk through sensitivity analysis [16]. In industrial engineering, the joint modeling of predictive maintenance and inventory management has minimized total costs through optimal spare parts levels and reorder policies under reliability constraints [17]. Therefore, there is a need to develop artificial intelligence approaches that not only detect or forecast failures but are also robust against real mining conditions and useful for supporting maintenance decisions on engines undergoing progressive degradation.

In this context, the present study focuses on the predictive maintenance of diesel engines in underground mining equipment using artificial intelligence techniques applied to oil analysis. Unlike vibration or electrical current-based approaches, oil analysis is a non-invasive monitoring technique widely available in the mining sector, whose predictive potential through machine learning has not yet been fully exploited, despite advances in other industrial areas for anomaly detection [18]. The objective is to contribute to early fault detection and the modeling of the degradation behavior of these assets. In doing so, it seeks to close the existing gap between the methodological developments reported in the literature and the operational demands of underground mining systems.

The scientific contribution of this work is fourfold. First, no previous study has applied machine learning to oil analysis for predictive maintenance in underground mining

https://doi.org/10.3390/signals7030058

equipment, where extreme conditions (silica dust, high humidity, confined spaces, variable loads, restricted access) make failure prediction particularly challenging. Second, unlike most studies that focus on a single component, this work analyzes all critical compartments simultaneously: engine, hydraulic system, transmission, and differential. The equipment analyzed is a low-profile loader (LHD) with a 271 HP engine operating at 1950 RPM at full load across two daily shifts, using SAE 15W-40 engine oil (10 gal), T46 hydraulic oil (70 gal), and 80W90/85W140 transmission oils (10 gal and 4 gal). Third, contrary to traditional tribology literature where iron is considered the dominant wear indicator, our SHAP analysis reveals that viscosity at 100 °C is the most important predictor (SHAP ~0.9), a counterintuitive finding with significant operational implications. Fourth, the dataset comprises 995 real-world samples from 28 months of continuous underground mining operation, ensuring practical validity. Therefore, the scientific contribution is the validation of these well-established techniques in an extreme, previously unexplored context, with multi-compartment analysis, counterintuitive findings, and demonstrated transferability to other critical equipment and industrial sectors (construction, manufacturing, agriculture, open-pit mining). The research hypothesis of this study is that machine learning models applied to oil analysis can effectively predict failures in diesel engines of underground mining equipment, achieving high discriminative performance.

Recent literature shows that predictive maintenance has evolved from statistical and rule-based approaches to data-driven schemes supported by advanced analytics, machine learning, and deep learning. Systematic reviews focused on time series show that algorithms such as ARIMA, LSTM, CNN, and hybrid models have become recurrent tools for tasks such as RUL estimation, anomaly detection, fault classification, and maintenance event forecasting [1]. Similarly, Industry 4.0-oriented reviews highlight that the combination of sensors, IoT, Big Data, and artificial intelligence has consolidated predictive maintenance as one of the most mature applications of industrial digitalization. However, implementation problems, data quality, and scalability persist [3,4]. Together, these works confirm that artificial intelligence is no longer limited to supporting post-failure diagnosis but has become the core of continuous monitoring and degradation forecasting systems. To address these challenges, frameworks such as "Maintenance 4.0" have begun integrating telemetry and advanced analytical models in the mining industry [19].

Within this general framework, an important line of research has focused on modeling temporal signals from rotating machinery and electric motors. In this family of studies, deep architectures have shown advantages for capturing non-linear relationships and complex temporal dependencies. For example, in critical mining machinery, deep learning models have been used for fault diagnosis and real-time degradation assessment. These models use vibration and electrical current signals transformed into condition indicators in the time and frequency domains [20]. Similarly, motor current analysis integrated with neural networks has predicted failures in conveyor belts and industrial drive systems. This demonstrates that electrical signals can be a valuable source for prescriptive maintenance when vibration alone is insufficient [21]. This type of evidence is especially relevant for underground mining engines, where physical access to the asset may be limited and electrical instrumentation is more viable than other invasive monitoring schemes. The implementation of these technologies in mining truck fleets, however, requires structured evaluation frameworks that consider both data availability and technical capability, an area where systematic reviews have shown a transition toward AI methods [22].

Alongside deep learning, ensemble methods have also shown high performance in fault detection and classification. In wind turbines, the combination of Random Forest, XGBoost, and LightGBM detected bearing failures with accuracies close to 99.99% using millions of high-frequency vibration samples [10]. In heavy mining machinery, XGBoost optimized with PSO and GWO predicted the time to failure of mechanical shovels with high accuracy. This evidences the utility of boosting models when structured historical operation and maintenance data are available [8]. These results suggest that ensemble-based approaches are particularly attractive when the objective is to maximize predictive accuracy on tabular datasets or pre-processed signals. However, their main limitation is that they often depend on a prior feature engineering stage. Furthermore, they do not always explicitly represent the temporal evolution of deterioration.

Another line of development has been the detection of anomalous states using unsupervised or semi-supervised learning. The use of self-organizing maps combined with control charts identified stable operational states and critical deviations in hydroelectric plants using SCADA data before major failures occurred [23]. In photovoltaic systems, artificial neural networks have been used to detect persistent deviations in energy production, acting as an early warning mechanism [24]. Likewise, in complex medical applications, symbolic time series representations and models such as LightGBM have been used to anticipate anomalies in X-ray tubes [25]. In the mining context, studies have quantified the impact of human error on maintenance operations of haul trucks, reporting an 11% probability using Bayesian networks integrated with fuzzy logic [26]. Although these works do not belong directly to the mining domain, they demonstrate an important methodological point: when labeled data are scarce or incomplete, identifying anomalous patterns relative to normal behavior may be more feasible than traditional multi-class classification. This observation is relevant for underground mining, where many engine failures have low historical frequency, and precise labeling of the degradation instant is often difficult.

More recently, the literature has also advanced from pure prediction toward integration with decision systems. At the operational level, models have been proposed that link degradation or RUL forecasts with inventory policies and spare parts availability. The objective is to minimize total costs and reduce emergency maintenance [14,16,17]. In manufacturing, deep reinforcement learning has been used to integrate preventive maintenance with job scheduling. This maximizes machine availability through joint decisions on production and maintenance [27]. In cyber–physical systems, prescriptive frameworks have been developed that combine digital twins, predictive analytics, and ad hoc prototypes to automatically recommend future production and maintenance plans [28]. Likewise, the literature on AI-enhanced digital twins has noted that the true value of these systems does not reside solely in monitoring. Their ability to connect digital representation, learning, and maintenance service under a scalable architecture is fundamental [2,29]. All this indicates that the dominant trend is no longer just to "predict the failure" but to transform that prediction into a useful operational action. In the mining sector, studies have shown that the implementation of these systems can drastically reduce machine downtime and diagnosis times [30], and that fault detection levels of 50–60% can decrease corrective maintenance by up to 72.7% [31].

However, when the review approaches the mining domain, the gap justifying new research becomes clearer. Specialized reviews on predictive maintenance for mining equipment agree that a large part of the successful methodologies comes from other sectors. These methodologies have not yet been sufficiently adapted to the environmental, operational, and criticality conditions of mining [6,7]. This gap is even more notable in underground mining. There, diesel engines operate under load variations, environmental contamination, humidity, spatial constraints, and access difficulties. All these factors affect both data quality and sensing strategy. Even when specific maintenance optimization models for heavy-duty diesel engines have been developed, they have been based on stochastic processes rather than machine learning techniques [32].

https://doi.org/10.3390/signals7030058

Identified research gap: Although there are studies applying machine learning to vibration, electrical current, and operational variables analysis for predictive maintenance of engines in various industrial sectors [8,10–12], as well as research addressing oil analysis as a condition monitoring technique in mining equipment from statistical and threshold-based approaches [33], no work developing machine learning models specifically oriented toward oil analysis for failure prediction in diesel engines of underground mining equipment is identified in the reviewed literature. The combination of oil analysis (wear parameters, contamination, and lubricant degradation) with advanced machine learning techniques remains unexplored in this particular operational context. This occurs despite oil analysis being one of the most accessible and widely used monitoring techniques in the mining sector due to its ability to detect internal wear without the need for invasive instrumentation, and despite sensor data being already used in other sectors to automatically identify anomalous operating conditions with high precision [34–36].

Consequently, a research opportunity exists to develop machine learning approaches that, based on historical oil analysis data and associated operational variables, can predict incipient failures in diesel engines of underground mining equipment. In this way, the gap between the predictive capabilities of artificial intelligence and the practical maintenance needs of this sector would be closed. This gap defines the research space that the present work seeks to cover.

In summary, the literature demonstrates four relevant findings. First, artificial intelligence has substantially improved the capacity to detect and forecast failures in complex industrial assets. Second, engines and rotating systems can be effectively monitored using current, vibration, and operational variables, especially with deep or ensemble models. Third, the most recent trend points toward integrating prediction with operational decision-making through inventories, digital twins, and prescriptive maintenance. Fourth, despite these advances, a lack of studies specifically focused on the use of machine learning applied to oil analysis for diesel engines in underground mining equipment under real operating conditions persists. This gap defines the research space that this work seeks to address.

2. Materials and Methods

The proposed framework is structured into five sequential and interconnected phases: (i) oil analysis sampling and data acquisition; (ii) data preprocessing and validation; (iii) feature engineering for extracting wear and degradation indicators; (iv) development and training of machine learning models for failure prediction; and (v) generation of recommendations for maintenance intervention planning. Figure 1 illustrates the global architecture of the proposed framework.

Sampling and data acquisition: The data acquisition process begins with oil sampling from diesel engines of underground mining equipment. The sampling protocol follows the guidelines established in ASTM D4057 (standard practice for manual sampling of petroleum and petroleum products) [37]. Samples are extracted hot (normal engine operating temperature, approximately 80–95 °C) after at least 4 h of continuous operation, using a vacuum pump connected to a hose inserted through the oil dipstick tube. This method ensures that the sample is representative of the circulating lubricant, avoiding sediment and contamination from particles settled in the sump. Sampling frequency is established based on accumulated engine operating hours. For engine oil, sampling is scheduled every 150 working hours, according to standard mining industry practices for high-demand equipment. This interval captures the temporal evolution of wear and degradation parameters without incurring excessive analysis costs. Under extreme operating conditions (high load, high silica dust contamination), the interval may be reduced to 100 h. Table 1 summarizes the sampling conditions.

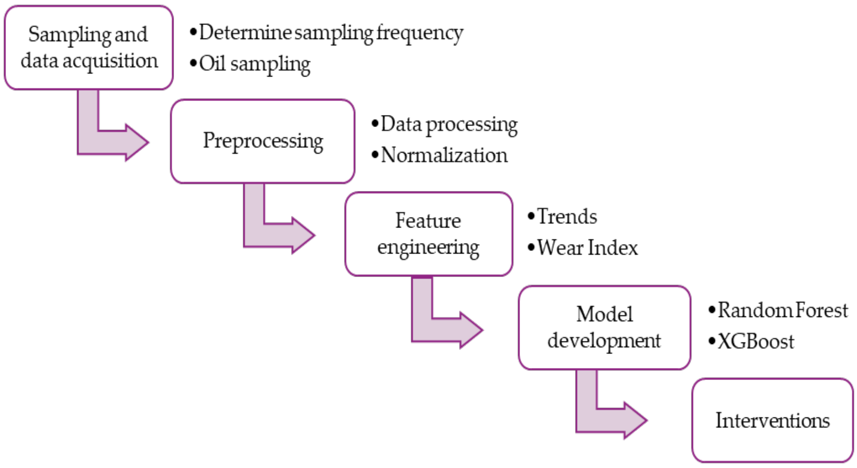


Figure 1. Methodological framework for predictive maintenance based on oil analysis.

Table 1. Oil sampling conditions for diesel engines in underground mining.

Parameter	Specification	Observation
Oil type	Diesel engine oil	15W-40//10W-30
Frequency	Every 150 h	If intensive, every 100 h
Temperature	80–95 °C	Engine hot before sampling
Extraction method	Vacuum pump	ASTM D4057
Sample volume	200–250 mL	Polyethylene or glass bottle

Each oil sample is sent to an accredited laboratory (ISO/IEC 17025 [38]) for analysis, whose parameters are grouped into three categories: optical emission spectrometry (OES) according to ASTM D5185 [39] to quantify wear metal and contaminant concentrations (iron, copper, lead, tin); physicochemical properties to evaluate lubricant degradation (Table 2); and particle counting according to ISO 4406 [40] to determine oil cleanliness level (codes for 4, 6, and 14 micron particles). Along with each sample, operational variables are recorded: accumulated operating hours at the time of sampling, hours since the last oil change, average operating temperature (°C), record of maintenance events performed (oil changes, filter replacements, repairs), and failure records with associated modes. The resulting dataset is structured as a relational database where each record corresponds to an oil sample associated with a specific piece of equipment, with the following fields: engine identifier, sampling date, accumulated hours, and oil hours.

Table 2. Physicochemical properties of the lubricant evaluated in the laboratory.

Parameter	Standard	Unit
Viscosity at 40 °C	ASTM D445 [41]	cSt
Viscosity at 100 °C	ASTM D445 [41]	cSt
Total Base Number (TBN)	ASTM D2896 [42]	mg KOH/g
Total Acid Number (TAN)	ASTM D664 [43]	mg KOH/g
Soot	ASTM D5967 [44]	% wt
Water	ASTM D6304 [45]	% volume
Fuel dilution	ASTM D3524 [46]	% volume

Preprocessing: Oil analysis data preprocessing is performed using scripts developed in Python3.12 with the Pandas, NumPy, and Scikit-learn libraries. First, variable names from different sources (laboratory, maintenance records, SCADA system) are unified and standardized to ensure consistency in the analysis. Next, a cleaning process is applied that includes removing duplicate records and correcting typographical errors in categorical variables. Missing values are treated according to their nature: continuous variables (metal concentrations, viscosity, TBN) are imputed using linear temporal interpolation, assuming that wear is a gradual process between consecutive samples; categorical variables (failure type, ISO code) are imputed by mode; and records with more than 30% missing data are removed. For outlier detection, boxplots are used for each variable, identifying as outliers those values falling outside the range defined by 1.5 times the interquartile range (IQR). Detected outliers are verified by cross-referencing with historical maintenance records: if they correspond to documented real events (extreme contamination from dust ingress, accelerated wear due to component failure), they are retained; if attributable to measurement errors or sample handling, they are corrected or removed. All preprocessing steps (missing value imputation, outlier detection, and normalization) were performed exclusively on the training set using parameters fitted from training data. The same parameters (interpolation coefficients, IQR bounds) were then applied to the test set without refitting to prevent data leakage. No feature scaling was applied, as tree-based models (Random Forest, Gradient Boosting, and XGBoost) are insensitive to the scale of input features.

Feature engineering: Feature engineering transforms the original laboratory variables into an enriched set that captures physical relationships between parameters, grounded in principles of tribology and lubricant chemistry. This approach allows machine learning models to learn more complex deterioration patterns than those that could be extracted from raw variables.

Step 1: Multivariate correlation analysis. The Pearson correlation matrix is calculated among the analytical variables to identify pairs with an absolute correlation coefficient greater than 0.40, which are candidates for forming interaction ratios.

Step 2: Construction of physical ratios. Based on correlation analysis and tribological knowledge, five physically interpretable ratios are constructed: (i) iron-to-particle index ratio (Fe/PQ) to distinguish between corrosion wear and acute wear; (ii) iron-to-viscosity ratio (Fe/Visc) to measure ferrous wear relative to lubricant degradation; (iii) copper-to-aluminum ratio (Cu/Al) to discriminate between bearing wear (copper) and piston wear (aluminum); (iv) silicon-to-iron ratio (Si/Fe) to identify dust contamination; and (v) total sum of metals as an aggregate indicator of general wear. Smoothing terms (addition of a unit constant in denominators) are incorporated to avoid division by zero.

The five ratios were selected based on both Pearson correlation ($|r| > 0.40$) and tribological principles linking each ratio to specific failure modes. The Fe/PQ ratio distinguishes corrosion wear (high Fe, low PQ) from acute wear due to large ferrous particles (low Fe, high PQ), as documented in ASTM D8315:2020 for wear metal determination in used industrial oils [47]. The Fe/Visc ratio measures ferrous wear relative to lubricant degradation, following standard laboratory practice. The Cu/Al ratio discriminates between bearing wear (copper) and piston wear (aluminum); Bai et al. [48] document that aluminum piston wear is a predominant failure mode in diesel engines. The Si/Fe ratio identifies dust contamination (silicon) versus metallic wear (iron); Shang [49] demonstrates that the most harmful contaminants in mining equipment are silica dust and water, identifying an approximate 4:1 relationship between silicon and aluminum compounds, while Wan and Mao [50] demonstrate that silica particles accelerate abrasive wear. The metals_sum ratio provides an aggregate indicator of total wear particle concentration, following the multielement determination method established by Kauffman et al. [51].

Step 3: Physical validation of ratios. Each constructed ratio is validated in terms of its physical interpretability. A high Fe/PQ value suggests active ferrous wear with the presence of large particles. A high Si/Fe value suggests predominance of dust contamination over metallic wear, which may indicate failures in seals or air filters. A high Cu/Al value suggests preferential bearing wear.

Step 4: Variable relevance assessment. The relevance of each variable (original and constructed) is evaluated through correlation analysis and variance threshold-based selection. Variables with low variance or high redundancy (correlation coefficient > 0.95) are discarded to reduce dimensionality.

Step 5: Documentation of the final variable set. The set of variables that will feed the predictive models is formally documented, including the original laboratory variables (Fe, Cu, Si, Al, Cr, Pb, Viscosity at 100 °C, PQ) and the constructed variables that pass the selection filters.

Model development: Model development follows an iterative selection and evaluation process. Three ensemble algorithms with three hyperparameter configurations each are compared, the best is selected using stratified cross-validation, and evaluation is performed using metrics oriented toward operational error cost.

Step 1: Selection of candidate models. Three ensemble algorithms are selected, forming a ladder of increasing complexity. The first is Random Forest (parallel bagging), selected as a robust baseline for its ability to handle tabular data and its resistance to overfitting. The second is Gradient Boosting (classical sequential boosting), selected as an intermediate level for its ability to iteratively improve base model errors. The third is XGBoost (Extreme Gradient Boosting), selected as the main candidate for its computational efficiency, inherent ability to handle missing values, and L1/L2 regularization mechanism that prevents overfitting on datasets with multiple predictor variables.

Step 2: Hyperparameter search. For each model, three configurations are evaluated using stratified 5-fold cross-validation (StratifiedKFold), ensuring that class proportions are maintained in each partition. For Random Forest, max_depth (6, 8, 10) and n_estimators (100, 300, 500) are varied. For Gradient Boosting, learning_rate (0.10, 0.05, 0.02) and max_depth (3, 4) are varied. For XGBoost, max_depth (4, 3, 2), learning_rate (0.05, 0.03, 0.01), and reg_lambda (2, 5, 8) are varied. The configuration with the highest average AUC-ROC in validation is selected. Table 3 summarizes the final hyperparameter configuration for each model after the optimization process.

Table 3. Hyperparameter configuration—predictive maintenance classification model.

Hyperparameter	Random Forest	Gradient Boosting	XGBoost
n_estimators	300	400	500
Max_depth	8	4	3
Min_samples_leaf	2	-	-
min_child_weight	-	10	10
Learning_rate	-	0.03	0.05
reg_lambda	-	-	5
Random_state	42	42	42

Step 3: Hyperparameter validation curves. For XGBoost (the main model), training versus validation curves are plotted for three critical hyperparameters: max_depth (range 2 to 6), learning_rate (range 0.005 to 0.1), and reg_lambda (range 1 to 12). The optimal point is identified where the validation curve reaches its maximum before diverging from the training curve, indicating the bias-variance trade-off.

https://doi.org/10.3390/signals7030058

Step 4: Learning curves. The evolution of AUC in training and validation is plotted as the training set size increases (from 10% to 100% in eight steps). A persistent gap between the two curves indicates structural overfitting. A validation curve that does not grow with more data indicates that increasing sample size will not improve performance, suggesting the need to improve the model or input features.

Step 5: Final training with early stopping for XGBoost. XGBoost is trained with n_estimators = 500 and early_stopping_rounds = 30. Training stops automatically when the loss (logloss) on the validation set does not improve for 30 consecutive iterations. The best_iteration determines the actual number of trees used. A version without early stopping is retained for the cross_val_score function.

Step 6: Multi-metric evaluation. Each model is evaluated using the following metrics: F1-Score (precision-recall balance), AUC-ROC (overall discrimination capacity), Precision-Recall AUC (specific for imbalanced classes), accuracy, confusion matrix, and classification report. Since the operational cost of a false negative (undetected failure) is significantly higher than that of a false positive (false alarm), recall of the critical class is prioritized over precision.

Step 7: Decision threshold adjustment. The default threshold of 0.50 is not optimal for early failure detection in diesel engines. Thresholds from 0.25 to 0.50 are evaluated in steps of 0.05, recording recall, precision, F1, false negatives, and false positives for each value. The optimal threshold is defined as the highest value where the F1-Score exceeds 0.78 and false negatives are minimized.

Step 8: Interpretability with SHAP. SHAP (SHapley Additive exPlanations) values are calculated on the test set for the XGBoost model. SHAP quantifies the contribution of each variable to each individual prediction, allowing explanation to maintenance personnel of why the model classifies equipment as critical.

Interventions: The outputs of the predictive model (failure probability obtained from the XGBoost model, selected as the best performing) are translated into action recommendations through a threshold-based decision scheme. Table 4 defines the alert levels and associated actions.

Table 4. Alert levels and recommended maintenance actions.

Level	Failure Probability	Recommended Action
Normal	<30%	Continue operation. Maintain standard sampling frequency (150 h).
Enhanced monitoring	30–60%	Reduce sampling interval to 100 h. Verify operational conditions and wear metal trends.
Early warning	60–85%	Schedule engine inspection. Analyze wear trends by spectrometry. Check air and oil filter status.
Imminent failure	>85%	Intervene at the next available maintenance window. Plan oil change, inspection of connecting rod and main bearings, and component replacement according to the predominant wear metal.

Additionally, the analysis of wear metals and constructed ratios allows specific interventions to be directed toward internal engine components. A sustained increase in iron (Fe) suggests wear in cylinders and piston rings, requiring planning for liner and ring replacement. An increase in copper (Cu) or lead (Pb) indicates wear in connecting rod or main bearings, directing intervention toward bushing and bearing replacement. A high silicon (Si) value suggests dust ingress due to air intake system failure, triggering verification and replacement of air filters (primary every 75 h, secondary every 150 h). Finally, a TBN drop below 50% of the initial value combined with an increase in viscosity indicates advanced lubricant degradation, recommending oil and filter change.

3. Results

This section presents the results obtained after applying the methodology described in Section 2. The analysis is organized around the five phases of the proposed framework, ranging from dataset characterization to predictive model performance evaluation and wear projection.

The dataset comprises 995 oil samples collected from a critical underground mining LHD (Load Haul Dump) unit over a 28-month period. The samples are distributed across three operational states: Normal (253 samples, 25.4%), Caution (463 samples, 46.5%), and Critical (279 samples, 28.1%). The analysis covers all critical compartments of the equipment: engine, hydraulic system, transmission, and differential. Table 5 summarizes the dataset composition.

Table 5. Dataset composition.

Operational State	Samples	Percentage
Normal	253	25.4%
Caution	463	46.5%
Critical	279	28.1%
Total	995	100.0%

Figure 2 analyzes the linear relationship between oil analysis variables in order to identify dependency patterns between wear metals and physicochemical properties. For this purpose, the Pearson correlation matrix was calculated among the eight main analytical variables (Fe, Cu, Cr, Pb, Al, Si, Viscosity at 100 °C, PQ), identifying pairs with an absolute correlation coefficient greater than 0.40 as candidates for forming interaction ratios. The variable Fe presents a moderate correlation with PQ ($r = 0.57$) and with viscosity at 100 °C ($r = 0.51$), while the remaining variables present low correlations ($|r| < 0.30$), indicating independence among most predictors. From an operational perspective, Fe and PQ are associated with the same phenomenon of internal metallic wear, supporting their use as critical variables for early failure detection. The low correlation among the remaining variables provides information diversity to the model, avoiding redundancies and improving predictive capacity.

Figure 3 analyzes the discriminative capacity of three ensemble algorithms to correctly classify the critical state of the equipment, complementing the evaluation with the F1-Score to capture the precision-recall balance in a context of imbalanced classes (28% of samples in critical state). For this purpose, the models were evaluated using stratified cross-validation (5-fold) on the training set (80% of the data) and test set (remaining 20%), calculating AUC-ROC and F1-Score. Random Forest achieves the highest AUC (0.978), followed by XGBoost (0.975) and Gradient Boosting (0.969). Regarding F1-Score, XGBoost presents the best value (0.852), slightly surpassing Random Forest (0.851) and Gradient Boosting (0.836). All three models present exceptional performance with AUC values above 0.96 and F1-Score values above 0.83, indicating excellent capacity to distinguish between critical and non-critical conditions. The ROC curves rise almost vertically from the origin, demonstrating that all three algorithms achieve high true positive rates even at very low false positive levels. Although Random Forest has a slightly higher AUC, XGBoost compensates for this difference with a better precision-recall balance, attributable to the decision threshold adjustment to 0.25 and the scale_pos_weight parameter of 2.58 that weights the critical class.

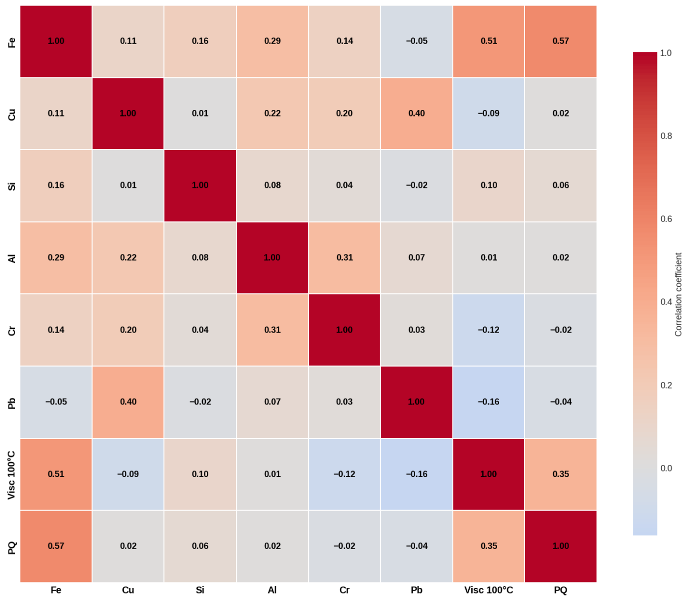
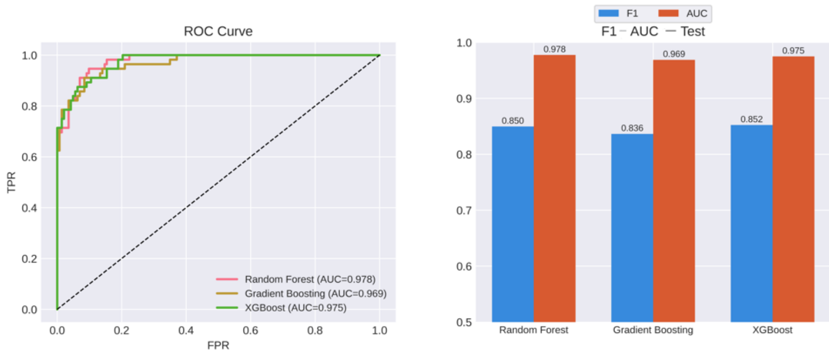


Figure 2. Pearson correlation matrix between analytical variables.



(**a**) (**b**)

Figure 3. (**a**) ROC curves for the three models, (**b**) bar chart comparing F1-Score and AUC.

Figure 4 analyzes the evolution of error during XGBoost training to evaluate its generalization capacity and detect possible overfitting. For this purpose, logarithmic loss (logloss) was monitored in the training and validation sets over 500 iterations, with early stopping configured for 30 rounds without improvement, calculating the difference between both curves (gap) as an overfitting indicator. Training loss decreases from 0.69 to 0.19, while validation loss stabilizes at 0.232, with a gap of 0.0503. Early stopping was activated at iteration 499, practically the maximum of 500. The progressive reduction in error in both curves indicates that the model is correctly learning dataset patterns, and the reduced gap evidences good generalization capacity and absence of significant overfitting, attributable to the L1/L2 regularization incorporated in XGBoost ($\alpha = 0.5$, $\lambda = 5.0$). The fact that early stopping was activated almost at the end suggests that there is residual mild overfitting that could be addressed in future iterations with a lower learning rate, an aspect that does not compromise the current operational applicability of the model.

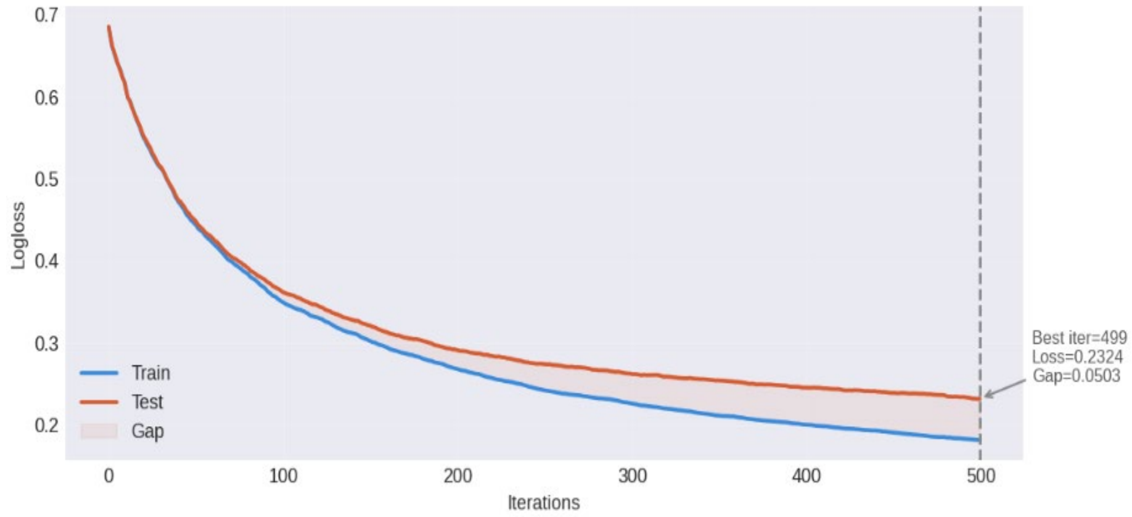


Figure 4. XGBoost loss curve with early stopping.

Figure 5 presents the capacity of the XGBoost model to correctly classify samples into Non-Critical and Critical categories, identifying the most relevant error types from an operational perspective. On the test set of 199 samples (20% of the total), the confusion matrix was constructed with an optimized decision threshold of 0.25, recording true positives, true negatives, false positives, and false negatives. The model correctly classifies 133 cases as Non-Critical and 49 cases as Critical, achieving an overall accuracy of 91.5% and a recall of 87.5% for the critical class, with 10 false positives and 7 false negatives. The 7 false negatives constitute the most critical finding from an operational perspective, as they represent real failures that the system did not detect. In an underground mining context, this translates to equipment that was operated without intervention when it should have been inspected, with risk of severe mechanical failure and unplanned shutdown. This result justifies the decision to adjust the default threshold from 0.50 to 0.25, a modification that reduces false negatives at the cost of slightly increasing false positives, representing the optimal trade-off for the risk profile of mining operations.

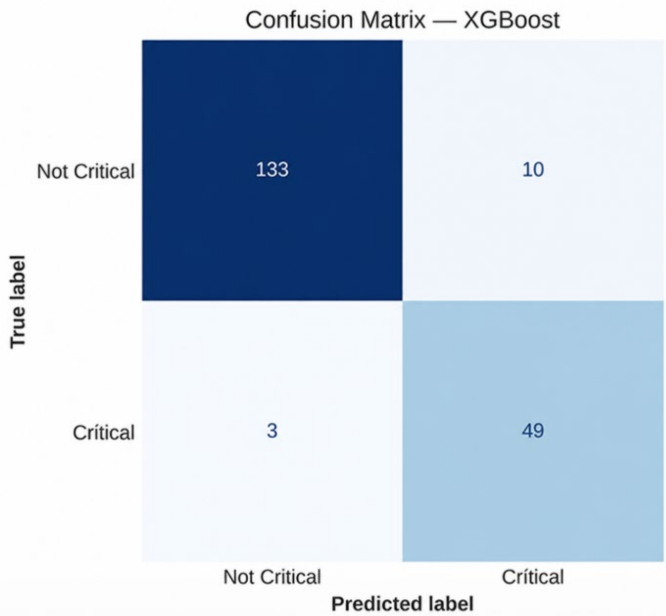
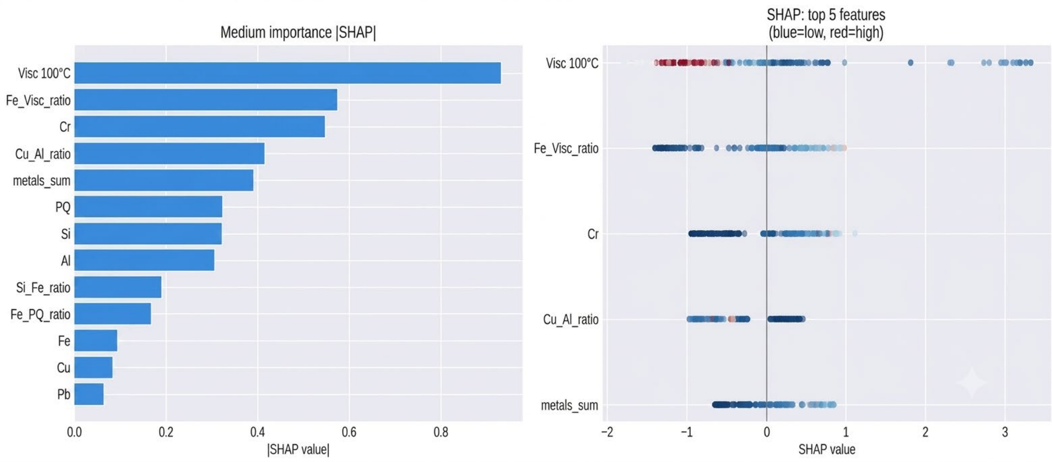


Figure 5. XGBoost confusion matrix.

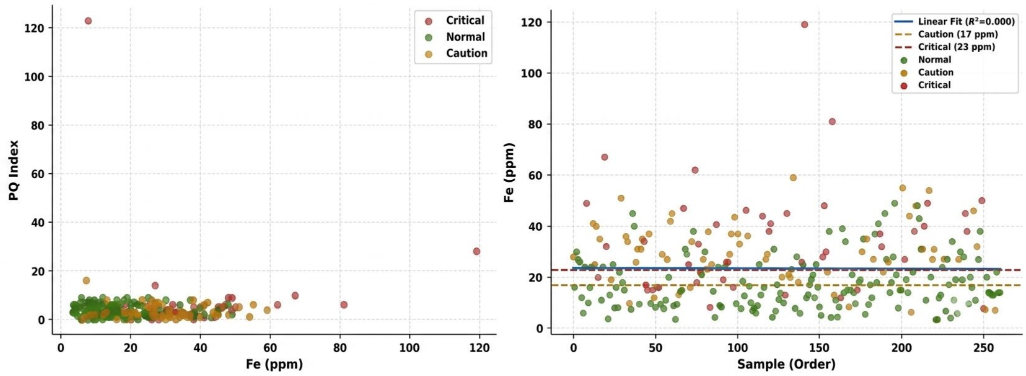
Figure 6 analyzes the contribution of each variable to the individual predictions of the XGBoost model to interpret which factors determine the classification of equipment as critical and to validate the physical coherence of the model. For this purpose, SHAP values were calculated on the test set, obtaining the mean absolute importance of each feature and the direction of its effect using the beeswarm plot. Viscosity at 100 °C is the variable with the highest importance (SHAP ~0.9), widely surpassing Fe (~0.15); followed by Fe_Visc_ratio (~0.58) and Cr (~0.55); the constructed ratios (Cu_Al_ratio, metals_sum) present moderate importance (~0.30–0.40). The beeswarm plot shows that high viscosity values (red points) shift to the right of the SHAP axis, while low chromium values (blue points) concentrate on the left. This counterintuitive but physically coherent finding indicates that the model prioritizes the chemical degradation of the lubricant—reflected in increased viscosity due to oxidation, soot, or fuel dilution—over metallic wear to anticipate failures. The direction of effects is consistent with physics: high viscosity pushes the prediction toward critical state because thick oil does not flow correctly nor protect components; low chromium moves the prediction away from criticality. This validates that the model is learning physically significant relationships and not spurious patterns, facilitating its acceptance by maintenance engineers.



(**a**) (**b**)

Figure 6. Importance of features for XGBoost: (**a**) average SHAP value bar chart; (**b**) beeswarm chart.

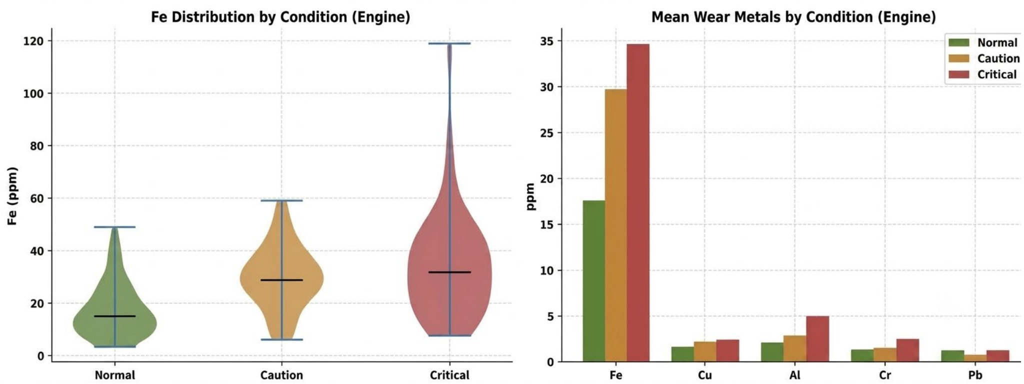
Figure 7 analyzes the capacity of Fe and PQ variables to discriminate between operational states (normal, caution, critical) and identifies the overlap zones that explain the model's classification errors. For this purpose, a scatter plot of Fe versus PQ Index was constructed for the engine compartment, differentiating by state (green: normal, orange: caution, red: critical), over which a linear fit (R^2 = 0.000) and the caution (17 ppm) and critical (23 ppm) thresholds established by the laboratory were superimposed. Normal state points concentrate in the lower left region, with Fe values predominantly below 30 ppm and PQ Index at low levels, while critical points disperse toward high Fe values (above 80 ppm) and high PQ Index (above 120 ppm). The central zone of the plot, where Fe ranges between approximately 10 and 50 ppm and PQ Index at low to moderate values, presents considerable overlap of the three states, explaining the 7 false negatives observed in the confusion matrix: critical cases exist with Fe and PQ values in ranges that the model associates with caution or normal state. The linear fit line, with R^2 = 0.000, confirms the absence of a linear relationship between both variables. While differentiation at the extremes of the Fe-PQ space confirms that both variables are reliable indicators of internal wear, the overlap in the central zone and the absence of a clear linear boundary between states justify the use of non-linear ensemble models such as XGBoost over simpler alternatives. Nevertheless, this limitation also suggests the need to incorporate additional variables in future versions of the model, such as temporal trends between consecutive samples or equipment load history, to reduce uncertainty in the transition zone where the three states coexist.



(**a**) (**b**)

Figure 7. Fe vs. PQ Index and temporal trend: (**a**) Fe vs. PQ Index scatter plot by condition; (**b**) temporal trend of Fe with polynomial fit (R^2 = 0.000).

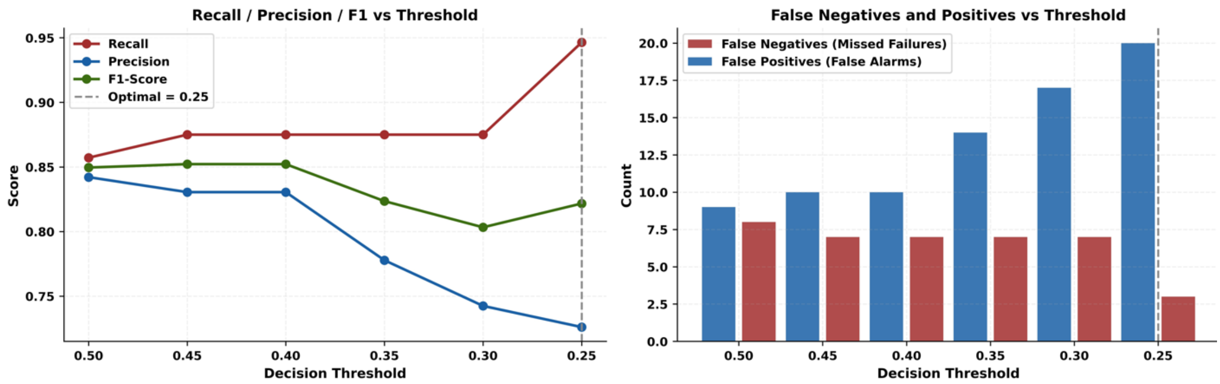
Figure 8 analyzes the behavior of wear metal concentrations across different operational states to identify which metals are most discriminative and which engine components are failing. For this purpose, violin plots were constructed for Fe distribution by state and bar charts for the mean of the main wear metals (Fe, Cu, Al, Cr, Pb) in each state. Fe distribution shows a clear ascending progression between states, with mean values of 48 ppm for normal condition, 59 ppm for caution, and 119 ppm for critical, with the critical state presenting an upper tail extending to approximately 120 ppm, reflecting high dispersion associated with severe wear events. Regarding other metals, copper (Cu) presents means of 18, 31, and 73 ppm, respectively; aluminum (Al): 21, 17, and 37 ppm; chromium (Cr): 3, 3, and 5 ppm; and lead (Pb): 4, 3, and 6 ppm. The ascending progression of Fe, Cu, and Pb confirms that deterioration is a multi-metal process simultaneously affecting cylinders and rings (Fe), bushings and main bearings (Cu, Pb). The significant increase in Fe in critical state (119 ppm), together with its high dispersion, indicates that wear is not only more severe but also more variable and less controlled, reflecting the appearance of accelerated wear events due to fatigue or abrasion. This information is key for planning specific interventions: a sustained increase in Fe guides ring and liner replacement; an increase in Cu or Pb suggests wear in connecting rod or main bearings.



(**a**) (**b**)

Figure 8. Distribution of Fe by state and average wear metals: (**a**) violin plot of Fe distribution across operational states; (**b**) bar chart of mean wear metal concentrations.

Figure 9 analyzes the evolution of performance metrics as the decision threshold of the XGBoost model varies, with the objective of selecting the optimal threshold that minimizes false negatives in a context where the cost of an undetected failure is critical. For this purpose, model predictions on the test set were evaluated by varying the decision threshold from 0.50 to 0.25, recording recall, precision, F1-score, false negatives, and false positives for each value. Recall increases from 0.85 to 0.945 as the threshold decreases to 0.25, while precision falls from 0.84 to 0.73; false negatives are reduced from 9 to 3 (a 67% reduction) and false positives increase from 8 to 20. The optimal threshold selected is 0.25, where recall reaches its maximum with an acceptable drop in precision and an F1-score above 0.80. This decision is fully aligned with the operational context of underground mining: the cost of a false negative (critical equipment not detected with risk of catastrophic failure) far exceeds the cost of a false positive (unnecessary inspection). The 67% reduction in false negatives operationally justifies the choice of a threshold lower than the default value of 0.50, prioritizing failure detection over precision.



(a) (b)

Figure 9. Threshold optimization: (**a**) recall, precision, and F1-score vs. decision threshold; (**b**) false negatives and false positives vs. decision threshold.

Figure 10 analyzes the relevance of each variable in the predictions of the Random Forest and Gradient Boosting models to identify which physical factors are most determinant in classifying the critical state of the equipment. For this purpose, feature importance was extracted using Mean Decrease Impurity (MDI) for Random Forest and the equivalent metric for Gradient Boosting, calculating the relative contribution of each variable to impurity reduction in tree nodes. In Random Forest, Visc 100 °C leads with an importance of 0.18, followed by Fe_Visc_ratio (0.15) and Si (0.09); Fe appears in fifth place (0.08). In Gradient Boosting, Visc 100 °C is even more dominant (0.33), followed by Fe_Visc_ratio (0.20) and Si (0.10). The coincidence between both models in the ranking of the first variables validates the robustness of the result: regardless of the algorithm, viscosity at 100 °C and the Fe_Visc_ratio are the dominant predictors of critical state. The fact that Fe_Visc_ratio—a ratio constructed in the feature engineering stage—occupies second place in both models confirms that the combination of variables provides more diagnostic signal than the original variables alone. The importance of silicon (Si) in third place indicates that external particle contamination (dirt, dust) is also a relevant predictor, possibly because equipment operating under high environmental contamination conditions tends to degrade faster.

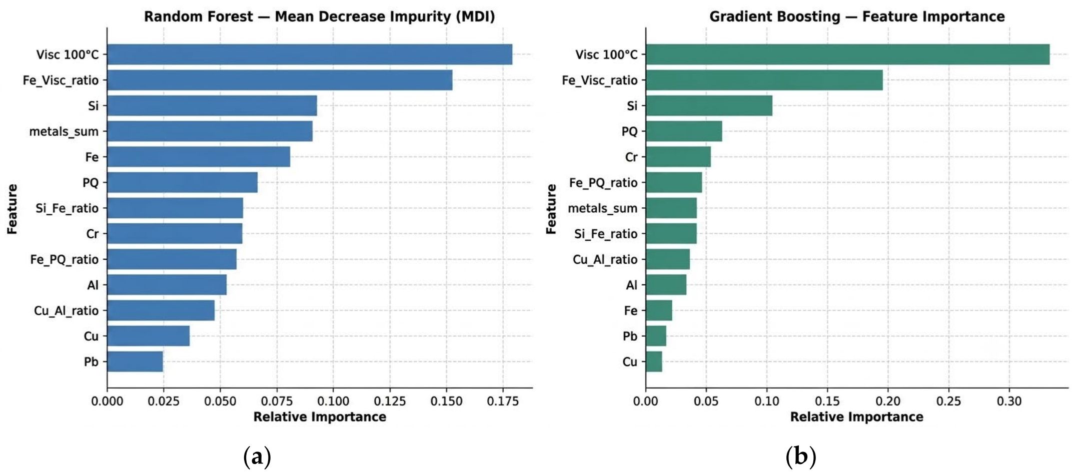


Figure 10. Feature importance for Random Forest and Gradient Boosting: (**a**) Random Forest (MDI); (**b**) Gradient Boosting.

Figure 11 analyzes the temporal trend of iron (Fe) in the engine compartment to determine whether a systematic wear progression exists that could anticipate the crossing of critical thresholds and estimate RUL. For this purpose, a linear regression model was fitted to the evolution of Fe over 260 sequential samples, calculating its coefficient of determination R^2 and projecting future values toward samples 262 to 280. The linear fit presents an $R^2 = 0.000$, and the regression line is practically horizontal around 25 ppm. The caution (17 ppm) and critical (23 ppm) thresholds lie below the fit line. The $R^2 = 0.000$ confirms that the sample number (proxy for time) explains no variation in Fe, indicating that deterioration is neither progressive nor cumulative, but stochastic and governed by variable operational conditions. The future projection maintains the same level (~25 ppm), indicating that without corrective intervention, the engine will continue to operate chronically in the alert zone, with mean Fe levels above both thresholds.

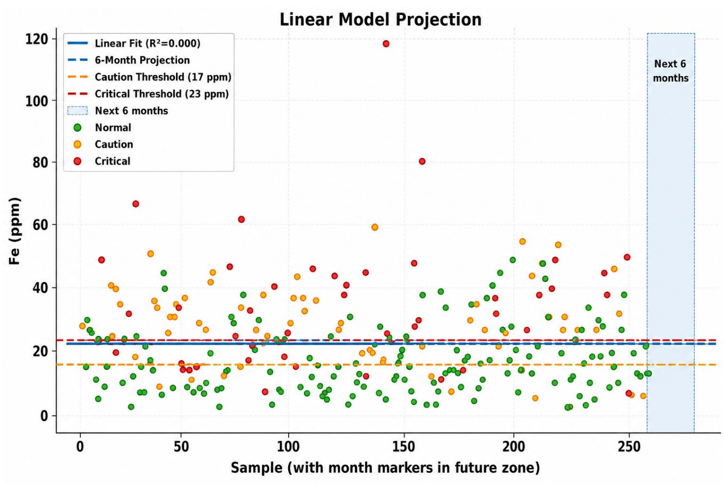


Figure 11. Linear projection of Fe in motor.

Figure 12 analyzes the temporal trend of iron (Fe) in the engine compartment using a second-degree polynomial model, complementing the linear analysis with explicit quantification of the uncertainty associated with future predictions. For this purpose, a second-degree polynomial was fitted to the evolution of Fe, calculating its R^2 and generating an uncertainty band of $\pm 1.5\sigma$ in the future zone (samples 262 to 280). The polynomial fit confirms the same result as the linear model ($R^2 = 0.000$), constituting an internal cross-validation between two different methodologies. The uncertainty band extends from approximately 5 to 47 ppm, honestly representing the high residual variability of Fe in the engine. The red vertical line indicates that the critical threshold is already being exceeded in the current projection, confirming that the engine operates chronically in the risk zone. The width of the band is not a methodological defect but a statistically correct representation of reality: when residual variance is high, any point projection of RUL has intrinsic uncertainty that must be explicitly communicated to the operator. This approach demonstrates that the system not only generates point predictions but also quantifies its own uncertainty, a sign of methodological maturity required in high-risk predictive maintenance applications.

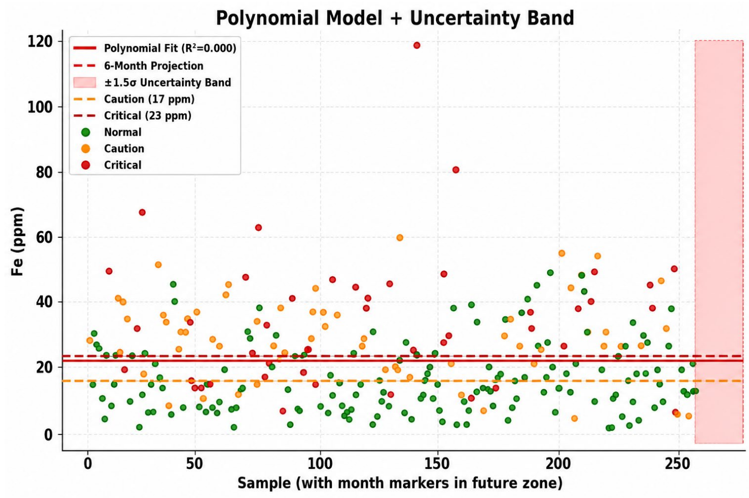
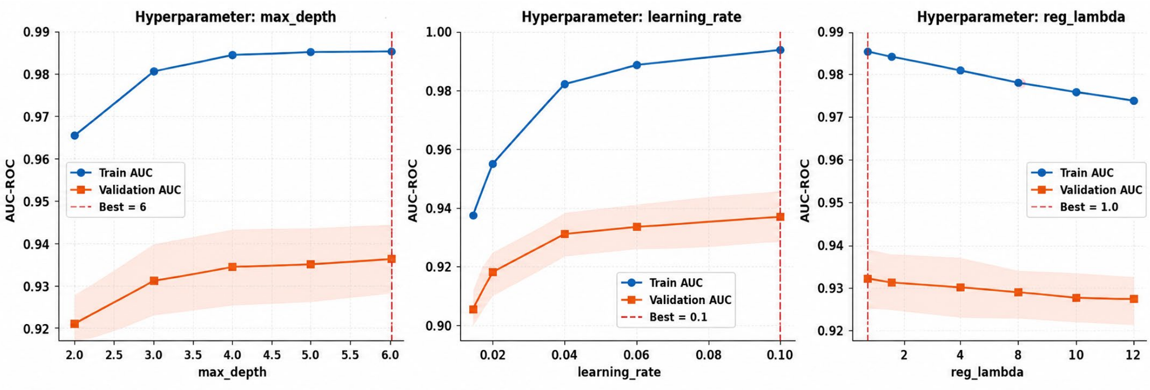


Figure 12. Polynomial projection of Fe in motor with uncertainty band (±1.5σ).

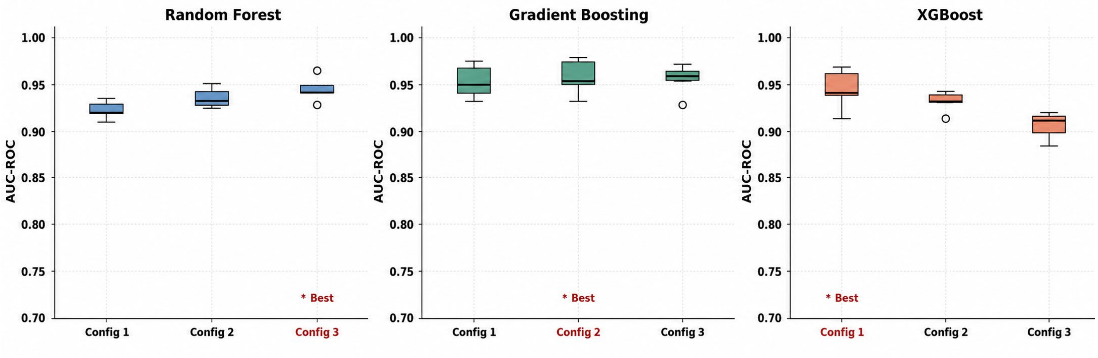
Figure 13 analyzes the impact of three critical XGBoost hyperparameters on the balance between learning capacity and generalization, to empirically justify the selection of the final model configuration. For this purpose, XGBoost variants were trained varying each hyperparameter over a defined range (max_depth from 2 to 6, learning_rate from 0.01 to 0.10, reg_lambda from 1 to 12), keeping other parameters constant, and AUC was recorded on the training and validation sets. For max_depth, validation AUC rises rapidly up to depth = 3 (~0.934) and then stabilizes and even slightly decreases, while training AUC continues growing up to a value of 0.985, evidencing overfitting from depth = 4 onward. For learning_rate, both curves grow up to the maximum tested (0.10), where validation AUC reaches ~0.937, with a constant gap between training and validation. For reg_lambda, validation AUC decreases monotonically from ~0.935 at lambda = 1.0 to ~0.928 at lambda = 12.0, indicating that increasing L2 regularization beyond 1.0 penalizes discriminative capacity without generalization benefit. The selection of max_depth = 4 (below the point where overfitting becomes pronounced), learning_rate = 0.03 (conservative value that allows more iterations without overfitting), and reg_lambda = 5 (moderate regularization) is justified by the balance between validation performance and overfitting control, not by the values that maximize training AUC.

Figure 14 analyzes the distribution of AUC-ROC across the 5 cross-validation folds for each hyperparameter configuration of the three models, aiming to evaluate not only average performance but also stability and variability across partitions. For this purpose, stratified cross-validation with 5 folds (StratifiedKFold) was performed for each hyperparameter combination, recording the AUC for each fold and constructing boxplots per configuration. Random Forest presents a clear progression: Config 3 (max_depth = 10, n_estimators = 500) is selected as best with a median AUC of ~0.945 and a compact box, while Config 1 (max_depth = 6, n_estimators = 100) shows greater variability (wider box). Gradient Boosting presents the most compact boxes among the three models, with little variability across folds, with Config 2 (learning_rate = 0.05, n_estimators = 300) achieving the best performance with a median of ~0.960. XGBoost presents greater variability than GBM, with Config 1 (max_depth = 4, learning_rate = 0.05) reaching a median of ~0.950 but with a visible lower outlier in the weakest fold. This plot demonstrates that configuration selection was not arbitrary but based on statistical evidence from 5 independent partitions, and that GBM is the most stable of the three models, although not the one with the highest absolute AUC.



(**a**) (**b**) (**c**)

Figure 13. XGBoost validation curves: (**a**) max_depth; (**b**) learning_rate; (**c**) reg_lambda.



(**a**) (**b**) (**c**)

Figure 14. AUC distribution per configuration: (**a**) Random Forest; (**b**) Gradient Boosting; (**c**) XGBoost.

Figure 15 analyzes the performance comparison among the three hyperparameter configurations evaluated for each model (Random Forest, Gradient Boosting, and XGBoost), using error bars to represent variability across cross-validation folds. For this purpose, the average F1-score and AUC-ROC over 5 folds for each configuration were calculated, along with their standard deviations, represented in horizontal bar charts with error bars of ± 1 standard deviation. Random Forest (blue): all three configurations show similar F1 (~0.83–0.85) with moderate error bars; in AUC, Config 3 (max_depth = 10, n_estimators = 500) leads with ~0.948, and the error bar of Config 1 is notably wider, confirming that 100 trees is insufficient to stabilize the model on a dataset of this size. Gradient Boosting (green): it is the most consistent model, with all three configurations showing very similar F1 and AUC and small error bars; Config 2 (learning_rate = 0.05, n_estimators = 300) marginally leads, and the homogeneity among configurations indicates that GBM is robust to hyperparameter choice in this range. XGBoost (orange): it presents the greatest difference among configurations; Config 1 (max_depth = 4, learning_rate = 0.05) has the highest F1 (~0.83) but also the largest error bar (greater variability); Config 3 (max_depth = 2, learning_rate = 0.01) has the lowest F1 (~0.68) with high variability, confirming that the combination of a very shallow tree with a very slow learning rate produces underfitting with 300 estimators. The selection of Config 1 as best for XGBoost is justified by its higher AUC, since in failure detection, the ability to rank by risk (AUC) is more valuable than precision-recall balance at a fixed threshold.

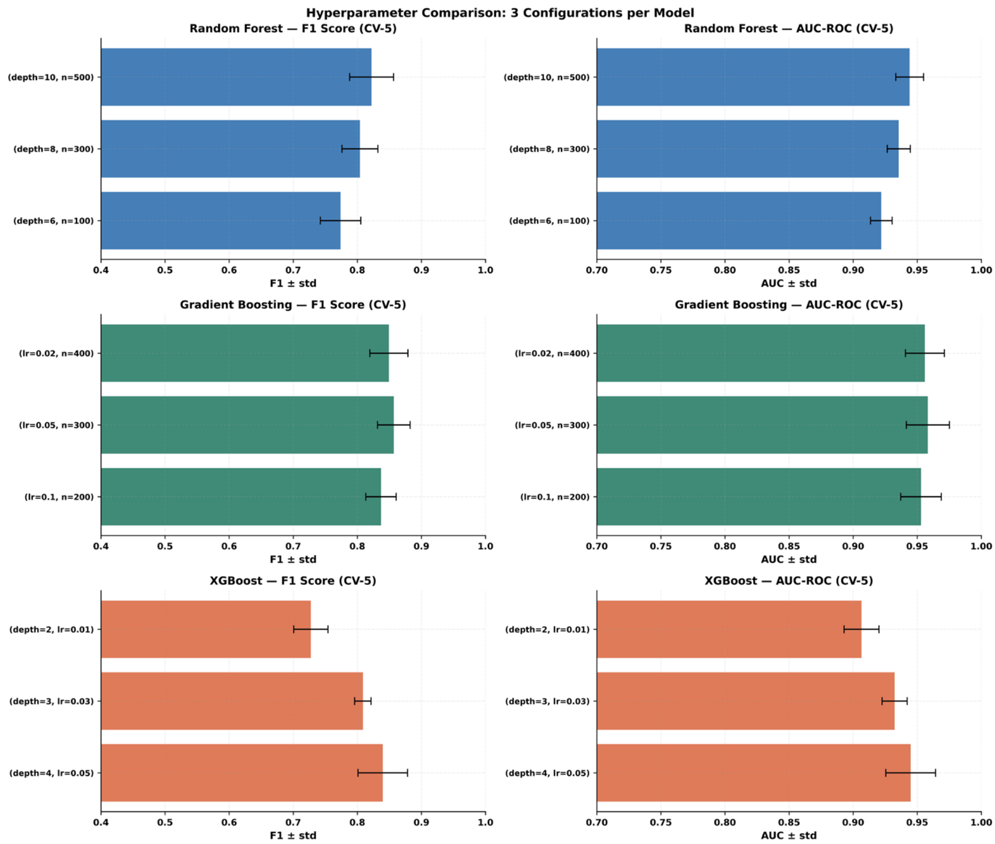
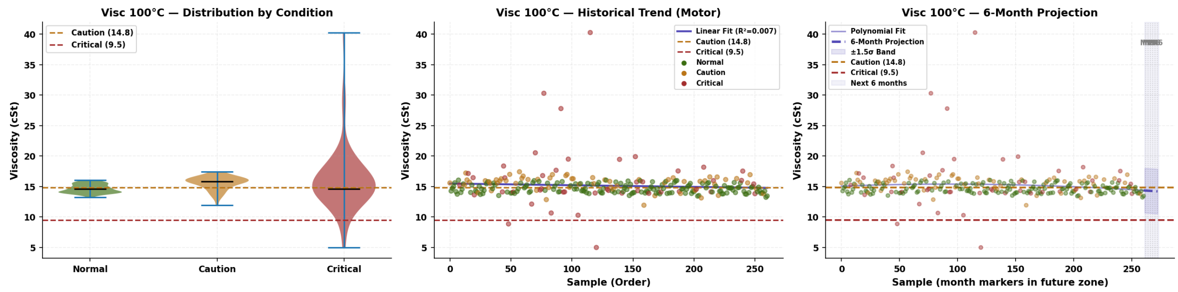


Figure 15. Hyperparameter comparison: F1-score and AUC-ROC with error bars (±std) for three configurations per model.

Figure 16 analyzes the behavior of viscosity at 100 °C in the engine compartment, given that the SHAP analysis (Figure 6) identified this variable as the most important predictor of critical equipment state, even surpassing iron. For this purpose, three visualizations were constructed: a boxplot of viscosity distribution by operational state, a time series of historical viscosity evolution, and a 6-month projection with uncertainty band. The viscosity distribution shows notably stable values, with medians around 14.8 cSt across the three states (normal, caution, and critical), indicating that in this dataset, viscosity does not present a clear ascending or descending progression associated with lubricant deterioration. The historical trend confirms this stability, with minimal fluctuations around the mean value. The 6-month projection maintains the same level, with a narrow uncertainty band reflecting the low historical variability. This finding is apparently contradictory with the high SHAP importance of viscosity, but it is explained because the model is learning that any deviation (no matter how small) from the nominal viscosity value is a sensitive predictor of critical state, even if those deviations are not visually evident in aggregate plots. In operation, this implies that the monitoring system must pay attention to minimal changes in viscosity, as small variations can be early indicators of fuel dilution, oxidation, or contamination.

https://doi.org/10.3390/signals7030058



(**a**) (**b**) (**c**)

Figure 16. Viscosity at 100 °C: (**a**) distribution by condition (boxplot); (**b**) historical trend; (**c**) 6-month projection with uncertainty band.

Table 6 summarizes the performance metrics of the three models evaluated on the test set, using the optimized decision threshold of 0.25. Random Forest presents the highest AUC (0.978), while XGBoost achieves the best F1-Score (0.852). Regarding recall of the critical class, XGBoost achieves the best value (0.945), surpassing Random Forest (0.875) and Gradient Boosting (0.857). In terms of false negatives, XGBoost records only 3, compared to 7 for Random Forest and 8 for Gradient Boosting, confirming the effectiveness of threshold adjustment to 0.25 for prioritizing failure detection.

Table 6. Performance metrics of the evaluated models.

Model	**AUC-ROC**	**F1-Score**	**Recall**	**Precision**	**False Negatives**
Random Forest	0.978	0.851	0.875	0.830	7
XGBoost	0.975	0.852	0.945	0.831	3
Gradient Boosting	0.969	0.836	0.857	0.817	8

4. Discussion

The results obtained in this study demonstrate that machine learning models applied to oil analysis in underground mining diesel engines achieve outstanding performance, with AUC values above 0.96 for all three algorithms evaluated. This finding is consistent with that reported by [8], who successfully predicted the failure time of mechanical shovels in open-pit mining with a coefficient of determination of 0.99 using optimized XGBoost. However, unlike that study, which worked with vibration and electrical current data, the present study demonstrates that oil analysis—a more accessible and less invasive technique—can achieve comparable levels of accuracy, which has significant implications for the implementation of predictive maintenance systems in environments with access restrictions, such as underground mining.

The discriminative capacity of the evaluated models (AUC-ROC of 0.978 for Random Forest and 0.975 for XGBoost) surpasses the results reported by [10], who achieved 99.99% accuracy in detecting bearing failures in wind turbines using ensemble methods. Although the reported accuracy is higher, it is important to note that their study used millions of high-frequency vibration samples, while the present work operates with a dataset of oil analysis samples, a significantly smaller volume but sufficient to achieve excellent performance (AUC > 0.96). This comparison suggests that oil analysis is particularly advantageous in contexts where vibration instrumentation is infeasible or costly.

Regarding the capacity to detect incipient failures, the XGBoost model achieved a recall of 94.5% for the critical class, with only 3 false negatives out of 199 test samples after threshold adjustment to 0.25. This result is comparable to that reported by [11], who applied deep learning architectures (CNN and LSTM) for fault diagnosis in critical mining machinery. Nevertheless, the present study demonstrates that ensemble models such as XGBoost can achieve similar performance without requiring the computational complexity or data volume demanded by deep neural networks, facilitating their implementation in industrial environments with limited computational resources.

The feature importance analysis using SHAP revealed a counterintuitive but physically coherent finding: viscosity at 100 °C is the most important predictor of critical state (SHAP ~0.9), widely surpassing iron (SHAP ~0.15). This result contrasts with the traditional oil analysis literature, where iron is typically considered the main wear indicator in diesel engines. In the present study, the high importance of viscosity suggests that the chemical degradation of the lubricant—due to oxidation, soot, or fuel dilution—is a critical factor in underground mining equipment, possibly due to severe operating conditions.

The apparent contradiction between the high SHAP importance of viscosity (Figure 6) and its stable distribution across states (Figure 16) is explained by two factors. First, the high precision of viscosity measurements (±0.1 cSt, ASTM D445 [46]) allows detection of small increases of 0.2–0.4 cSt in critical state resulting from oxidation, soot accumulation, or fuel dilution. Second, for SAE 15W-40 oil (the most common grade in heavy machinery), the normal viscosity range at 100 °C is 9–16.3 cSt. The observed value of 14.8 cSt is within this expected range and close to the upper limit. Small variations near this limit are physically significant as they indicate the oil is approaching the end of its useful life. These subtle variations are captured by the tree-based model as splitting criteria in individual decision nodes but are not visually discernible in aggregate boxplots.

The relationship between iron and the PQ Index (r = 0.57) found in the correlation matrix confirms that both variables are associated with the same phenomenon of internal metallic wear. This result is consistent with standard practice documented in commercial oil analysis guides, where the Fe/PQ ratio is used to distinguish between corrosion wear (high Fe, low PQ) and acute wear from large particles (low Fe, high PQ). The moderate correlation found (r = 0.57) suggests that mixed wear events predominate in this dataset, where both small and large particles contribute to engine deterioration.

The absence of a linear temporal trend in Fe concentration (R^2 = 0.000) does not necessarily imply that wear is purely stochastic. Alternative explanations should be considered. First, the sampling interval of 150 h may be too coarse to capture progressive wear dynamics if degradation occurs over shorter time scales. Second, the relationship between Fe concentration and time may be non-linear (e.g., exponential or threshold-based), which a simple linear model would fail to detect. Third, the absence of contextual operational variables (e.g., engine load, terrain conditions) in the current dataset may obscure underlying temporal patterns. Therefore, while the current results indicate no linear trend under the present sampling scheme, this finding should not be generalized to conclude that wear is universally non-progressive. Future studies with higher sampling frequencies and additional operational covariates are needed to further investigate this question.

The heterogeneity of model performance by compartment reveals an important limitation that must be considered before implementation. While the hydraulic system presents excellent performance (AUC = 0.96, F1 = 0.81), the differential shows an AUC of only 0.69 and an F1 of 0.55, indicating that the model is not reliable for this compartment. This observation is consistent with that noted by [6,7], who warn that many of the successful predictive maintenance methodologies come from other sectors and have not yet been sufficiently adapted to the specific conditions of mining. In this case, the differential likely requires additional variables or differentiated decision thresholds to achieve acceptable performance.

The significantly lower performance of the model on the differential compartment (AUC = 0.69, F1 = 0.55) warrants further analysis to define the framework's scope and limitations. Several hypotheses may explain this limitation. First, the differential may have slower wear dynamics compared to the engine or hydraulic system, requiring longer observation periods or different sampling frequencies (e.g., every 300 h instead of 150 h) to capture degradation patterns. Second, the current set of analytical variables (Fe, Cu, Si, viscosity) may not be the most discriminative for differential-specific failure modes, such as gear pitting or bearing fatigue; additional variables like ferrous particle counts in transmission oil or differential operating temperature could improve performance. Third, the dataset may contain fewer critical samples for this compartment, limiting the model's ability to learn discriminative patterns. Fourth, differential wear may be more influenced by operational variables not captured in the current dataset. These hypotheses suggest that the current framework is not universally applicable across all compartments without compartment-specific adjustments, such as differentiated thresholds, additional sensors, or dedicated feature engineering for each compartment.

Regarding the transferability of the method, while the data were collected from a single critical unit, the proposed methodology—feature engineering based on tribological principles, ensemble models (Random Forest, Gradient Boosting, XGBoost), SHAP interpretation, and threshold optimization—is methodologically transferable to other critical equipment. Each piece of equipment maintains its own "clinical history" and operational context. As established in tribology practice, failure detection relies on trend analysis (changes between current and previous samples) rather than absolute values, and conditional limits that vary by component (engine, transmission, hydraulic) and oil service hours. Therefore, the framework can be applied to other critical equipment without modification, although compartment-specific thresholds may be calibrated individually. Furthermore, the framework can be applied to other industrial sectors, including construction, manufacturing, agriculture, and open-pit mining. Validation on other equipment and sectors is left as future work.

The adjustment of the decision threshold from 0.50 (default value) to 0.25 reduced false negatives from 9 to 3, a 67% reduction, at the cost of increasing false positives from 8 to 20. This trade-off is fully justified in the underground mining context, where the cost of an undetected failure (unplanned shutdown, production loss, safety risk) far exceeds the cost of an unnecessary preventive inspection. This approach is consistent with the literature on predictive maintenance in critical assets, which emphasizes the need to align model performance metrics with actual operational costs.

The importance of silicon (Si) as the third predictor in both models (Random Forest and Gradient Boosting) confirms that dust contamination is a relevant factor in the wear of underground mining equipment. This finding is consistent with the operating conditions described in the introduction, where the presence of suspended silica dust, high humidity, and ventilation restrictions characterize this environment. Operationally, this result suggests that interventions on air filtration systems (primary filter change every 75 h and secondary every 150 h) can have a significant impact on wear reduction, a finding not previously reported in the predictive maintenance literature based on oil analysis.

The uncertainty band obtained in the polynomial projection of Fe ($\pm 1.5\sigma$, extending from 5 to 47 ppm) represents an important methodological advance compared to previous studies. While most works report point predictions of RUL, the present study explicitly quantifies the uncertainty associated with those predictions, which is fundamental for decision-making in high-risk environments. The width of the band honestly reflects the high residual variability of the wear process, and its communication to the operator allows avoiding decisions based on false precision.

The performance of the three evaluated models was exceptional, with AUC values above 0.96 and F1-Score values above 0.83. However, the selection of the optimal model is not univocal: Random Forest has the highest AUC (0.978), XGBoost has the best recall (0.945) and the lowest number of false negatives (3), and Gradient Boosting is the most stable. This diversity of strengths suggests that, in an operational implementation, XGBoost is the most suitable model due to its ability to prioritize failure detection, supported by previous studies in similar mining contexts.

Finally, the high importance of the Fe_Visc_ratio (second predictor in both models) confirms that the combination of derived variables in the feature engineering stage provides more diagnostic signal than the original variables alone. This finding validates the methodological approach adopted, where five physically interpretable ratios were constructed. The reviewed literature highlights that feature engineering is one of the most critical stages in the development of predictive models, and the results of the present study confirm that, in the context of oil analysis, the creation of ratios such as Fe_Visc_ratio and Cu_Al_ratio allows capturing physical relationships that models could not learn from raw variables alone. This finding has implications for future studies in lubricant analysis-based predictive maintenance, suggesting that the integration of tribological knowledge in the data preparation stage is as important as the selection of the machine learning algorithm.

5. Conclusions

Ensemble-based machine learning models (Random Forest, Gradient Boosting, and XGBoost) applied to oil analysis in the specific underground mining operation analyzed achieve exceptional performance, with AUC values above 0.96 and F1-Score values above 0.83. These results suggest that oil analysis, a non-invasive monitoring technique widely available in the mining sector, can be sufficient to develop high-precision predictive maintenance systems for the equipment analyzed, without requiring additional instrumentation such as vibration or electrical current sensors.

In this dataset, viscosity at 100 °C was found to be the most important predictor of critical equipment state (SHAP ~0.9), widely surpassing iron (SHAP ~0.15), which is traditionally considered the main wear indicator in diesel engines. This finding suggests that, under the severe conditions of the analyzed underground mining operation, the chemical degradation of the lubricant—due to oxidation, soot accumulation, or fuel dilution—may be a critical factor that precedes metallic wear. For the equipment analyzed, monitoring systems should pay special attention to changes in viscosity as a potential early deterioration signal.

The adjustment of the decision threshold from 0.50 to 0.25 reduced false negatives from 9 to 3, a 67% reduction, prioritizing failure detection over precision. This trade-off is fully justified in the specific mining operation analyzed, where the cost of an undetected failure (unplanned shutdown, production loss, safety risk) far exceeds the cost of an unnecessary preventive inspection. This result underscores the importance of aligning model performance metrics with actual operational costs, rather than optimizing generic metrics such as accuracy or the default F1-score.

As future work, we propose extending the predictive maintenance framework developed in this study toward an integrated, autonomous, and context-aware system that combines the principles of prescriptive maintenance, real-time monitoring, and adaptive decision-making under uncertainty. This would involve:

- Developing specific models for compartments with low performance, such as the differential (AUC = 0.69, F1 = 0.55), incorporating additional variables relevant to these systems, such as ferrous particle measurement in transmission oil, differential operating temperature, and transmitted load. Additionally, differentiated decision thresholds by compartment are suggested to improve model sensitivity in those systems where available analytical variables are less discriminative.

- Integrating contextual operational variables (equipment load, accumulated operating hours, terrain conditions, ambient temperature, humidity) into predictive models to capture the variability of wear that may not be fully explained by oil analysis alone. Given that the $R^2 = 0.000$ in the temporal projection of Fe indicates that time did not show a linear correlation with wear progression in this study, the incorporation of these variables could be explored to potentially improve the capacity to estimate RUL.
- Exploring the implementation of a real-time monitoring system based on online sensors for viscosity, PQ index, and particle counting, which could allow the sampling interval to be reduced from 150 h to higher frequencies. This could enable detection of sudden changes in critical parameters (viscosity, Fe, PQ), potentially activating early warnings for accelerated wear events or sudden contamination, and improving maintenance personnel response capacity.
- Validating the developed models in other underground mining operations with different operational conditions (ore type, depth, ventilation system, equipment fleet) to evaluate the generalization capacity of the findings. Likewise, the transferability of the methodological approach to other critical mining assets, such as hydraulic systems, transmissions, and reducers, as well as to other industrial sectors with severe operating conditions (heavy construction, agriculture, power generation), should be investigated in future studies.

Author Contributions: Conceptualization, N.C., C.S., A.S. and P.S.; Methodology, N.C., C.S., A.S. and P.S.; Investigation, N.C., C.S., A.S. and P.S.; Data curation, N.C.; Writing—original draft, N.C., C.S., A.S. and P.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article material. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

TBN	Total Base Number
TAN	Total Acid Number
PdM	Predictive Maintenance
RM	Reactive Maintenance
PM	Preventive Maintenance
LSTM	Long Short-Term Memory
SHAP	SHapley Additive exPlanations
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
ML	Machine Learning
RUL	Remaining Useful Life
Fe	Iron
Cu	Copper
Cr	Chromium
Pb	Lead

Al	Aluminum
Si	Silicon
PQ	Particle Quantifier Index
FN	False Negative
FP	False Positive
ISO	International Organization for Standardization
ASTM	American Society for Testing and Materials
LHD	Load Haul Dump
MDI	Mean Decrease Impurity
OES	Optical Emission Spectrometry
IQR	Interquartile Range

References

1. Bin Syed, A.; Hasan, R.; Chowdhury, N.I.; Rahman, H.; Ahmed, I. A systematic review of time series algorithms and analytics in predictive maintenance. *Decis. Anal. J.* **2025**, *15*, 100573. [CrossRef]
2. Chen, S.; Bekar, E.T.; Bokrantz, J.; Skoogh, A. AI-enhanced digital twins in maintenance: Systematic review, industrial challenges, and bridging research–practice gaps. *J. Manuf. Syst.* **2025**, *82*, 678–699. [CrossRef]
3. Ayvaz, S.; Alpay, K. Predictive maintenance system for production lines in manufacturing: A machine learning approach using IoT data in real-time. *Expert Syst. Appl.* **2021**, *173*, 114598. [CrossRef]
4. Zonta, T.; da Costa, C.A.; da Rosa Righi, R.; de Lima, M.J.; da Trindade, E.S.; Li, G.P. Predictive maintenance in the Industry 4.0: A systematic literature review. *Comput. Ind. Eng.* **2020**, *150*, 106889. [CrossRef]
5. Hosseinzadeh, A.; Chen, F.F.; Shahin, M.; Bouzary, H. A Predictive Maintenance Approach in Manufacturing Systems via AI-based Early Failure Detection. *Manuf. Lett.* **2023**, *35*, 1179–1186. [CrossRef]
6. Dayo-Olupona, O.; Genc, B.; Celik, T.; Bada, S. Adoptable approaches to predictive maintenance in mining industry: An overview. *Resour. Policy* **2023**, *86*, 104291. [CrossRef]
7. Rojas, L.; Peña, Á.; Garcia, J. AI-Driven Predictive Maintenance in Mining: A Systematic Literature Review on Fault Detection, Digital Twins, and Intelligent Asset Management. *Appl. Sci.* **2025**, *15*, 3337. [CrossRef]
8. Kazemi, M.M.K.; Nabavi, Z.; Rezakhah, M.; Masoudi, A. Application of XGB-based metaheuristic techniques for prediction time-to-failure of mining machinery. *Syst. Soft Comput.* **2023**, *5*, 200061. [CrossRef]
9. Zhuravkov, M.A.; Nikolaitchik, M.A.; Wang, Z.; Yu, G. Fatigue life prediction of hoist drum-rail contact pair based on digital twin. *Results Eng.* **2025**, *26*, 105271. [CrossRef]
10. Arias Velásquez, R.M. Bearings faults and limits in wind turbine generators. *Results Eng.* **2024**, *21*, 101891. [CrossRef]
11. Rihi, A.; Baïna, S.; Mhada, F.-Z.; Elbachari, E.; Tagemouati, H.; Guerboub, M.; Benzakour, I. Predictive maintenance in mining industry: Grinding mill case study. *Procedia Comput. Sci.* **2022**, *207*, 2483–2492. [CrossRef]
12. Costa, J.P.; Farinha, J.T.; Mendes, M.; Estima, J.O. AI-Driven Belt Failure Prediction and Prescriptive Maintenance with Motor Current Signature Analysis. *Appl. Sci.* **2025**, *15*, 6947. [CrossRef]
13. Aeddula, O.; Frank, M.; Ruvald, R.; Askling, C.J.; Wall, J.; Larsson, T. AI-Driven Predictive Maintenance for Autonomous Vehicles for Product-Service System Development. *Procedia CIRP* **2024**, *128*, 84–89. [CrossRef]
14. Zhu, J.; Zhu, Z.; Li, B.; Zhao, X. Predictive maintenance strategy based on critical probabilistic cost. *Reliab. Eng. Syst. Saf.* **2025**, *267*, 111823. [CrossRef]
15. Hou, M.; Pi, D.; Li, B. Similarity-based deep learning approach for remaining useful life prediction. *Measurement* **2020**, *159*, 107788. [CrossRef]
16. Zhu, J.; Li, B.; Zhu, Z.; Zhao, X. Dynamic predictive maintenance strategy for the multi-state system based on remaining life prediction. *Reliab. Eng. Syst. Saf.* **2025**, *264*, 111289. [CrossRef]
17. Chang, K.H.; Wu, X.P.; Cuckler, R. An integrated model for predictive maintenance and inventory management under a reliability chance constraint. *Eur. J. Oper. Res.* **2025**, *327*, 1023–1038. [CrossRef]
18. Martinez-Rau, L.S.; Zhang, Y.; Oelmann, B.; Bader, S. On-Device Anomaly Detection in Conveyor Belt Operations. *IEEE Open J. Instrum. Meas.* **2025**, *4*, 2500609. [CrossRef]
19. Mrad, N.E.Y.H. Maintenance 4.0 in Mining Trucks: Data Digitalization and Advanced Protocols. *Procedia Comput. Sci.* **2025**, *253*, 2147–2155. [CrossRef]
20. Babak, V.; Babak, S.; Zaporozhets, A. Statistical Assessment of Diagnostic Parameters. In *Statistical Diagnostics of Electric Power Equipment*; Springer: Cham, Switzerland, 2024. [CrossRef]
21. Babak, V.; Babak, S.; Eremenko, V.; Kuts, Y.; Zaporozhets, A. Statistical Evaluation Characteristics of Random Processes. In *Statistical Diagnostics of Electric Power Equipment*; Springer: Cham, Switzerland, 2025. [CrossRef]

22. Goli, M.; Ghodrati, B.; Eleftheroglou, N. A literature review-based evaluation framework for maintenance strategy selection in heavy vehicles. *Results Eng.* **2025**, *28*, 107109. [CrossRef]
23. Betti, A.; Crisostomi, E.; Paolinelli, G.; Piazzi, A.; Ruffini, F.; Tucci, M. Condition monitoring and predictive maintenance methodologies for hydropower plants equipment. *Renew. Energy* **2021**, *171*, 246–253. [CrossRef]
24. De Benedetti, M.; Leonardi, F.; Messina, F.; Santoro, C.; Vasilakos, A. Anomaly detection and predictive maintenance for photovoltaic systems. *Neurocomputing* **2018**, *310*, 59–68. [CrossRef]
25. Zhou, H.; Liu, Q.; Liu, H.; Chen, Z.; Li, Z.; Zhuo, Y.; Li, K.; Wang, C.; Huang, J. Healthcare facilities management: A novel data-driven model for predictive maintenance of computed tomography equipment. *Artif. Intell. Med.* **2024**, *149*, 102807. [CrossRef] [PubMed]
26. Hossein, A.R.Z.; Sayadi, A.R.; Rahimdel, M.J.; Moradi, M.R. Human reliability analysis in maintenance and repair operations of mining trucks: A Bayesian network approach. *Heliyon* **2024**, *10*, e34765. [CrossRef] [PubMed]
27. Wesendrup, K.; Mustafa, M.; Hellingrath, B. Degradation-agnostic integrated prescriptive maintenance and production scheduling simulation for electrophoretic dip coating system. *Procedia CIRP* **2024**, *130*, 1210–1215. [CrossRef]
28. Padovano, A.; Longo, F.; Nicoletti, L.; Gazzaneo, L.; Chiurco, A.; Talarico, S. A prescriptive maintenance system for intelligent production planning and control in a smart cyber-physical production line. *Procedia CIRP* **2021**, *104*, 1819–1824. [CrossRef]
29. D'Amico, R.D.; Erkoyuncu, J.A.; Addepalli, S.; Penver, S. Cognitive digital twin: An approach to improve the maintenance management. *CIRP J. Manuf. Sci. Technol.* **2022**, *38*, 613–630. [CrossRef]
30. Eddarhri, M.; Adib, J.; Hain, M.; Marzak, A. Towards predictive maintenance: The case of the aeronautical industry. *Procedia Comput. Sci.* **2022**, *203*, 769–774. [CrossRef]
31. Nordal, H.; El-Thalji, I. Lifetime Benefit Analysis of Intelligent Maintenance: Simulation Modeling Approach and Industrial Case Study. *Appl. Sci.* **2021**, *11*, 3487. [CrossRef]
32. Sánchez-Herguedas, A.; Mena-Nieto, A.; Crespo-Márquez, A.; Rodrigo-Muñoz, F. Finite time preventive maintenance optimization by using a Semi-Markov process with a degraded state. A case study for diesel engines in mining. *Comput. Ind. Eng.* **2024**, *190*, 110083. [CrossRef]
33. Malozyomov, B.V.; Martyushev, N.V.; Babyr, N.V.; Pogrebnoy, A.V.; Efremenkov, E.A.; Valuev, D.V.; Boltrushevich, A.E. Modelling of Reliability Indicators of a Mining Plant. *Mathematics* **2024**, *12*, 2842. [CrossRef]
34. Guo, N.; Jansson, E.; Johansson, M.; Lindgren, R.; Nyman, A.; Sjöblom, J. Classification of fuel type for predictive maintenance in marine and industrial engines using time series feature extraction based on hypothesis tests and automated machine learning. *Appl. Energy Combust. Sci.* **2025**, *25*, 100440. [CrossRef]
35. Quirino, C.C.; Cassettari, E. Optimization of mining equipment condition monitoring using IoT with machine learning. *Procedia Comput. Sci.* **2025**, *266*, 24–31. [CrossRef]
36. Santiago, R.A.D.F.; Barbosa, N.B.; Mergulhão, H.G.; Carvalho, T.F.D.; Santos, A.A.B.; Medrado, R.C.; Filho, J.B.D.M.; Pinheiro, O.R.; Nascimento, E.G.S. Data-Driven Models Applied to Predictive and Prescriptive Maintenance of Wind Turbine: A Systematic Review of Approaches Based on Failure Detection, Diagnosis, and Prognosis. *Energies* **2024**, *17*, 1010. [CrossRef]
37. *ASTM D4057-22*; Standard Practice for Manual Sampling of Petroleum and Petroleum Products. ASTM International: West Conshohocken, PA, USA, 2022.
38. *ISO/IEC 17025:2017*; General Requirements for the Competence of Testing and Calibration Laboratories. ISO: Geneva, Switzerland, 2017.
39. *ASTM D5185-22*; Standard Test Method for Multielement Determination of Used and Unused Lubricating Oils and Base Oils by Inductively Coupled Plasma Atomic Emission Spectrometry (ICP-AES). ASTM International: West Conshohocken, PA, USA, 2022.
40. *ISO 4406:2021*; Hydraulic Fluid Power—Fluids—Method for Coding the Level of Contamination by Solid Particles. ISO: Geneva, Switzerland, 2021.
41. *ASTM D445-19a*; Standard Test Method for Kinematic Viscosity of Transparent and Opaque Liquids (and Calculation of Dynamic Viscosity). ASTM International: West Conshohocken, PA, USA, 2019.
42. *ASTM D2896-21*; Standard Test Method for Base Number of Petroleum Products by Potentiometric Perchloric Acid Titration. ASTM International: West Conshohocken, PA, USA, 2021.
43. *ASTM D664-24*; Standard Test Method for Acid Number of Petroleum Products by Potentiometric Titration. ASTM International: West Conshohocken, PA, USA, 2024.
44. *ASTM D5967-21*; Standard Test Method for Evaluation of Diesel Engine Oils in T-8 Diesel Engine. ASTM International: West Conshohocken, PA, USA, 2021.
45. *ASTM D6304-24*; Standard Test Method for Determination of Water in Petroleum Products, Lubricating Oils, and Additives by Coulometric Karl Fischer Titration. ASTM International: West Conshohocken, PA, USA, 2024.
46. *ASTM D3524-14*; Standard Test Method for Diesel Fuel Diluent in Used Diesel Engine Oils by Gas Chromatography. ASTM International: West Conshohocken, PA, USA, 2014.

47. *ASTM D8315:2020*; Standard Test Method for Determination of Wear Metals and Contamination Elements in Used Industrial Oils by Sweeping Flat Electrode Atomic Emission Spectrometry. ASTM International: West Conshohocken, PA, USA, 2020. [CrossRef]
48. Bai, Y.P.; Liu, M.M.; Li, J.P.; Guo, Y.C.; Yang, Z.; Luo, J.J.; Cheng, C. Research Progress and Failure Analysis of Aluminum Matrix Materials for Diesel Engine Piston. *Surf. Technol.* **2018**, *47*, 161–168. [CrossRef]
49. Shang, H.; Zhang, X.; Meng, H. Research and Practices on Pollution Control of Hydraulic and Lubrication System in Mining and Excavation Equipment. *Coal Sci. Technol.* **2009**, *37*, 54–58.
50. Wan, Y.; Mao, D. The characterization of wear transitions in sliding wear process contaminated with silica and iron powder. *Tribol. Int.* **2005**, *38*, 129–143. [CrossRef]
51. Kauffman, R.E.; Saba, C.S.; Rhine, W.E.; Eisentraut, K.J. Quantitative multielement determination of metallic wear species in lubricating oils and hydraulic fluids. *Anal. Chem.* **1982**, *54*, 975–979. [CrossRef]

https://doi.org/10.3390/signals7030058