International Journal of Safety and Security Engineering
Vol. 16, No. 10, October, 2026, pp. 2275-2287
Journal homepage: http://iieta.org/journals/ijsse

Validation Strategy Determines Reported Performance in Machine Learning Models for Pump Condition Monitoring: Implications for Predictive Maintenance in Mining

Juan Christiann Samir Haro Zavaleta, Edinson Esmith Bocanegra Serin, Eduardo Manuel Noriega Vidal*, Walter Klysman Laiza Vera

Department of Mining Engineering, Faculty of Engineering, Universidad Privada del Norte, Trujillo 13001, Peru

Corresponding Author Email: eduardo.noriega@upn.edu.pe

Copyright: ©2026 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/ijsse.161007

Received: 17 August 2026
Revised: 29 September 2026
Accepted: 9 October 2026
Available online: 31 October 2026

Keywords:
Decision Tree, Gradient Boosting, industrial pumps, machine learning, predictive maintenance, Random Forest, Support Vector Machine, underground mining

ABSTRACT

Machine learning studies on pump condition monitoring often report near-perfect metrics, yet the validation designs behind them are rarely scrutinized. This study examines how the cross-validation strategy alone changes the reported performance of Random Forest (RF), Gradient Boosting (GB), Support Vector Machine (SVM) and Decision Tree (DT) models trained on a public industrial pump sensor dataset of minute-level observations. A single sample of 10,970 records was evaluated with randomly shuffled stratified k-fold and with chronological forward-chaining validation, with standardization, feature selection and hyperparameter tuning nested inside each training fold. Under stratified k-fold, all models reached F1-Scores between 0.9935 and 0.9966. Under chronological validation, performance depended on the degradation history available for training: with only 37 degraded records in the training window every model failed ($F1 \leq 0.003$), whereas with 781 or more, F1 ranged from 0.937 to 1.000. Regression of a Suction Pressure Index (SPI) showed the same pattern, with root mean squared error (RMSE) rising from 2.02-2.66 to 4.84-12.23 percentage points. No algorithm was consistently superior. Random cross-validation therefore masks the dependence of these models on prior failure episodes, and chronological validation is necessary before claiming predictive capability for maintenance decisions in mining.

1. INTRODUCTION

Underground mining represents one of the most demanding and critical operational environments from the standpoint of safety and production continuity. Among the auxiliary systems required to sustain underground operations, drainage pumps constitute one of the most critical components of the auxiliary services infrastructure, playing an essential role by controlling groundwater and preventing flooding in tunnels and working areas [1]. In the context of underground mining auxiliary services—which encompass ventilation, drainage, compressed air, and electrical systems—pump failure represents one of the highest-impact risks, as it can directly compromise the continuity of mining operations and the safety of underground personnel. Unexpected pump failures can lead to serious consequences: flooding of workings, production stoppages, risk to the lives of workers, and high recovery costs.

According to recent studies, pump system failures account for between 20% and 30% of unplanned downtime in underground mining operations [2]. These challenges highlight the need for predictive maintenance strategies capable of identifying performance deterioration before failure occurs. Continuous pump monitoring generates large volumes of sensor data, including variables such as pressure, temperature, vibration, flow rate, and electrical consumption. The complex relationships among these variables and the actual pump condition make traditional analytical approaches insufficient, especially in scenarios with nonlinear relationships and imbalanced classes [3].

In this context, machine learning has emerged as a robust approach for processing large volumes of multivariate data and accurately classifying critical operational states [4]. Previous studies have demonstrated the potential of machine learning in industrial pumping systems. The k-nearest neighbors (kNN) algorithm has been applied to detect hydraulic anomalies in pumping systems, achieving high accuracy in identifying failure patterns [5]. Decision Trees (DTs) have been used to classify failures in industrial pumps, with interpretability highlighted as a key advantage for engineering applications [6]. A machine learning framework for predictive maintenance of centrifugal pumps has also shown that ensemble approaches outperform single-model methods in health state assessment [7].

Among machine learning algorithms, Random Forest (RF) has shown particularly strong performance in industrial predictive maintenance applications. An RF-based model using pressure, flow, and motor current signals has achieved superior performance compared with other methods for multiclass fault classification in centrifugal pump systems [8]. RF has also been validated for predicting the lifespan of electric submersible pumps, demonstrating its practical applicability in industrial pump environments [9]. In

underground mining environments, recent research has also highlighted the effectiveness of machine learning techniques for improving operational reliability and safety under harsh operating conditions [10]. Additionally, ensemble methods such as Gradient Boosting (GB) and XGBoost have demonstrated competitive performance in industrial fault prediction tasks, achieving high accuracies in imbalanced sensor datasets [11, 12].

The validation methodology also plays a critical role in model reliability. Stratified k-fold cross-validation is particularly important for datasets with class imbalance, as it ensures each fold maintains the same class distribution as the original dataset, providing more reliable estimates of model performance [13, 14]. This technique has been shown to reduce variability in performance measurements and improve the consistency of results across different machine learning algorithms [15].

However, stratification addresses class imbalance but not temporal dependence. Pump sensor records are time series in which consecutive observations are strongly correlated, and when individual records are assigned to folds at random, observations from the same operating episode can appear in both training and test partitions. The resulting metrics may then reflect the recognition of episodes already seen during training rather than the ability to anticipate degradation in future operating periods. To our knowledge, the extent to which the validation strategy alone changes the performance reported for machine learning models in pump condition monitoring has not been examined systematically, although the question is directly relevant to predictive maintenance in underground mining, where a model that appears reliable in validation but fails in operation could leave a flooding risk undetected.

In this context, the present study evaluates four machine learning models for pump condition classification and for estimation of a Suction Pressure Index (SPI), and examines how the validation strategy affects the performance they appear to achieve. The specific objectives are: (1) to analyze the operational characteristics of pumps using industrial sensor data; (2) to implement RF, GB, Support Vector Machine (SVM) and DT for binary classification of operational states and for regression of the SPI; (3) to evaluate the models under randomly shuffled stratified k-fold and under chronological forward-chaining validation on the same sample, with feature selection and hyperparameter tuning nested inside each training fold; (4) to compare the models and the two validation strategies using classification and regression metrics; and (5) to identify the conditions under which the models can be expected to detect degradation in future operating periods.

2. METHODOLOGY

2.1 Data source

This study was based on the public Pump Sensor Data dataset available on the Kaggle platform [16], which contains records from industrial pump sensors in continuous operation. This dataset is widely used in predictive maintenance research and contains time-series measurements from 52 operational sensors captured at regular one-minute intervals over a seven-month monitoring period, documenting three operational states: NORMAL (correct operation), RECOVERING (progressive degradation), and BROKEN (failure). The sensors record physical parameters including suction and discharge pressure, fluid temperature, vibration, flow rate, and electrical consumption. It is important to note that drainage pumps used in underground mining auxiliary services operate under the same physical and mechanical principles as industrial centrifugal pumps: both are governed by the same hydraulic laws, are expected to exhibit broadly similar degradation patterns in pressure, temperature, vibration, and flow rate, and to respond to operational stress in comparable ways [4]. This physical analogy motivates the use of industrial pump sensor datasets as a starting point for developing predictive maintenance models. This strategy has precedent in the mining literature: Shao and Kumral [17] apply a public experimental dataset from a hydraulic test rig to predictive maintenance of mining machinery, including a case study on estimating the remaining lifespan of a water pump, on the grounds that the underlying machinery is representative of mining operations. However, direct empirical validation of these models in an underground mining environment has not yet been performed and is identified as a limitation in Section 5.1.

This physical analogy is grounded in the fact that both industrial centrifugal pumps and mining drainage pumps are governed by comparable failure physics: progressive wear of impellers and seals manifests as declining suction pressure, cavitation produces characteristic vibration signatures, and bearing deterioration raises operating temperature regardless of the installation environment. Consequently, the sensor-level degradation patterns captured in the Pump Sensor Dataset-the relationships among pressure, temperature, vibration, and flow rate that precede failure-are expected to be broadly analogous to those in underground mining drainage systems, where similar components fail through similar mechanisms-a hypothesis this study does not directly test with mining-site data.

2.2 Data processing and validation

Data review, filtering, and validation procedures were carried out before model training. The original dataset comprised 220,320 records. Sensor_15 was excluded because it contained no valid readings, and rows with a missing value in any of the remaining 51 sensor columns were then removed, leaving 119,103 valid records; no duplicate-removal step was applied. The RECOVERING and BROKEN states were grouped into a single category (Class 1: Degraded), since both represent operating conditions outside the optimal range and the BROKEN records numbered only seven instances, which were statistically insufficient for independent class modeling. Consequently, this binary formulation cannot distinguish gradual degradation from complete failure, and the results should not be interpreted as evidence of reliable identification of critical pump failures; they support flagging of degraded operating states in general.

To ensure computational efficiency while retaining every degraded observation, a sample of 10,970 records was drawn from the 119,103 valid records: all 2,970 degraded records and a random selection of 8,000 optimal records (random_state = 42). The degraded class is therefore deliberately over-represented relative to the full cleaned dataset, in which it accounts for about 2.5% of records. The resulting class balance is presented in Table 1. This single sample, ordered by timestamp, is the basis of every result reported in this study; no other sample was used.

Table 1. Balance of operational classes

Operational Status	Class	Sample Records
Optimal (NORMAL)	Class 0	8,000 (72.9%)
Degraded (RECOVERING + BROKEN)	Class 1	2,970 (27.1%)
Total		10,970 (100%)

Note: All 2,970 degraded records and 8,000 randomly selected optimal records from the 119,103 valid records; the degraded class is over-represented relative to the full cleaned dataset (about 2.5%).

2.3 Data preparation for modeling

To ensure scale consistency, input features were standardized with StandardScaler (zero mean, unit variance). Standardization and feature selection were implemented as steps of a single scikit-learn pipeline and were therefore refitted on the training partition of every fold, so that no information from a test partition influenced either the scaling parameters or the choice of predictors. Seven variables were selected at each fold (SelectKBest, k = 7), separately for each task. For classification, the candidates were the 51 available sensors, ranked by the analysis of variance (ANOVA) *F*-statistic with respect to the class label. For regression, sensor_00 was excluded from the candidates because the SPI is derived from it, and the remaining 50 sensors were ranked by the univariate F-statistic of their linear association with the SPI. Table 2 lists the variables selected in each fold. Within each fold the selection was identical for the four models, and it was highly stable across folds: in classification, the same seven variables were selected in all five folds of each strategy, the two strategies differing only in sensor_50 (stratified) and sensor_06 (chronological); in regression, the same seven variables were selected in all five stratified folds, while in chronological folds 1 to 4 sensor_51 was selected in place of sensor_50.

Table 2. Input variables selected in each outer fold, by task, and model outputs

Variable	Identifier	Classification	Regression
Suction pressure	sensor_00	All folds	Excluded (SPI derived from it)
Operational parameter 4	sensor_04	All folds	All folds
Flow rate	sensor_05	All folds	All folds
Operational parameter 3	sensor_06	Chronological folds	All folds
Fluid temperature	sensor_10	All folds	All folds
Operational parameter 1	sensor_11	All folds	All folds
Operational parameter 2	sensor_12	All folds	All folds
Vibration	sensor_50	Stratified folds	Stratified folds; chronological fold 5
Operational parameter 5	sensor_51	Not selected	Chronological folds 1-4
Operational status (output)	machine_status	Target	-
Suction Pressure Index (SPI), % (output)	min-max of sensor_00	-	Target

Note: Variables selected by SelectKBest (k = 7) inside each training fold. 'All folds' denotes the five outer folds of both validation strategies; 'Stratified folds' and 'Chronological folds' denote the five folds of that strategy only. Candidates: 51 sensors for classification and 50 for regression (sensor_00 excluded).

2.4 Operational class definition

The operational condition of each record was taken directly from the machine_status label supplied with the original dataset; no thresholds were computed by the authors from the predictor variables. This is important because deriving the target from the input features would render the classification task circular. A sample was assigned to Class 1 (Degraded) when its original label was RECOVERING or BROKEN, and to Class 0 (Optimal) when its original label was NORMAL, as detailed in Table 3. The strong separation between classes observed in the exploratory analysis is therefore an empirical property of the labeled data, not a consequence of the labeling rule: the two classes overlap in the extreme tails of suction pressure, yet none of the Class 1 records falls within the interquartile range of Class 0.

Table 3. Classification of operational states of pumps

Class	State	Description
Class 0	Optimal	Original dataset label: NORMAL
Class 1	Degraded	Original dataset label: RECOVERING or BROKEN

2.5 Machine learning

Machine learning is a branch of artificial intelligence that enables computer systems to learn from data without relying on explicitly programmed rules. In supervised learning, the objective is to develop a model capable of predicting outcomes for new, unseen data using datasets where both inputs and expected outputs are known. Classification and regression techniques are used to build effective predictive models that generalize beyond the training set [18].

2.5.1 Random Forest algorithm

RF is an ensemble learning method that combines multiple DTs to improve predictive accuracy and reduce variance through bagging and bootstrap aggregation [19]. In classification, the final prediction is determined by majority vote among all trees (Figure 1). The mean prediction in regression is expressed as:

$$\hat{y}(x) = (1/B) \times \sum_{b=1}^{B} T_b(x) \quad (1)$$

where, B is the total number of trees and $T_b(x)$ is the prediction of the b-th tree for input x. RF was configured with n_estimators = 100, class_weight = 'balanced', max_features = 'sqrt' and random_state = 42; max_depth was tuned.

2.5.2 Gradient Boosting algorithm

GB is a sequential ensemble method that builds models iteratively, where each new model corrects the residual errors of the previous one by minimizing a differentiable loss function [20]. Each iteration adds a new weak learner fitted to the negative gradient of the loss.

$$F_m(x) = F_{m-1}(x) + \eta h_m(x) \quad (2)$$

where, $F_m(x)$ is the model at iteration m, η is the learning rate, and $h_m(x)$ is the weak learner fitted to the residuals. GB was configured with n_estimators = 100, learning_rate = 0.1, max_features = 'sqrt' and random_state = 42; max_depth was tuned.

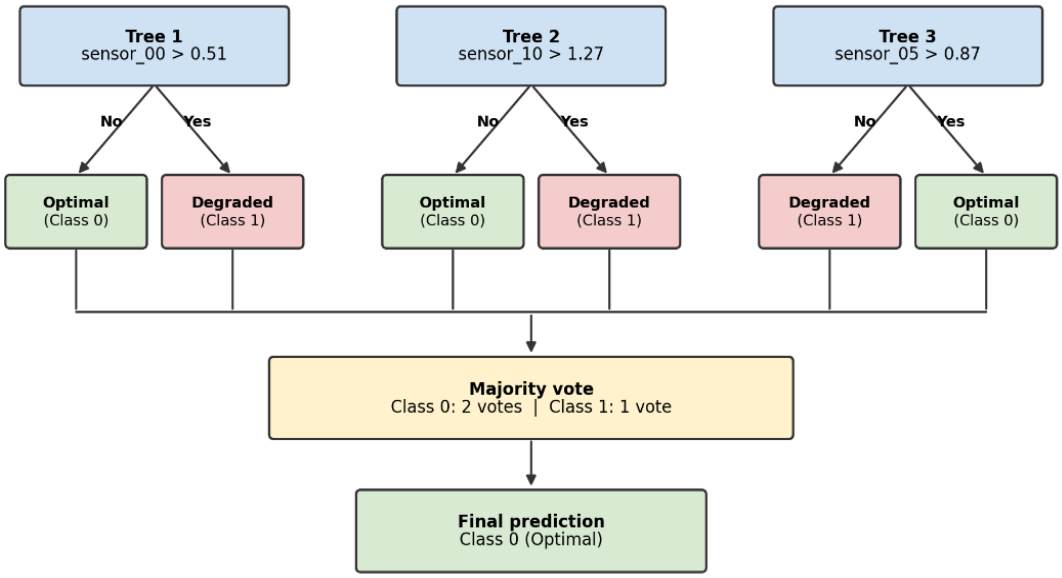


Figure 1. Random Forest (RF)-ensemble of decision trees with majority vote

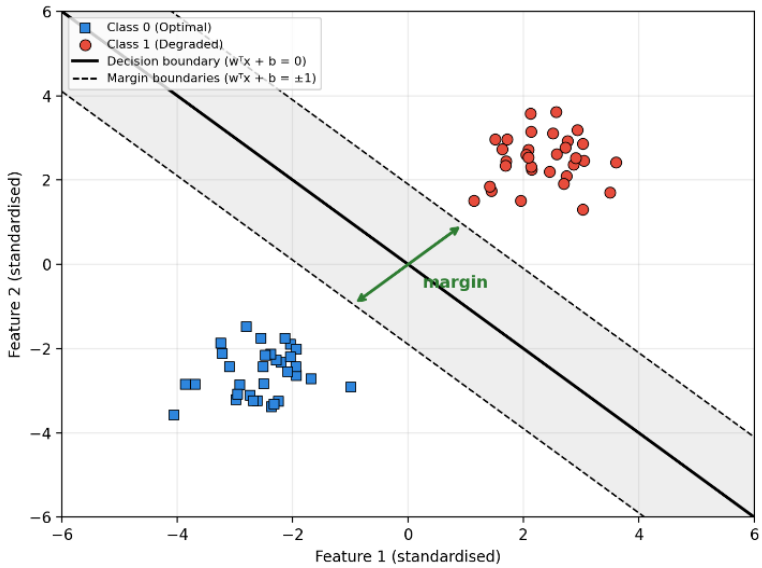


Figure 2. Support Vector Machine (SVM) decision boundary and maximum margin hyperplane

2.5.3 Support Vector Machine algorithm

SVM identifies the optimal hyperplane that maximizes the margin between classes (Figure 2). The kernel-based decision function is expressed as [21]:

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b \quad (3)$$

where, α_i are Lagrange multipliers, y_i are class labels, $K(x_i, x)$ is the Radial Basis Function (RBF) kernel defined as $K(x_i, x) = \exp(-\gamma \|x_i - x\|^2)$, and b is the bias term. SVM was configured with kernel = 'rbf', gamma = 'scale' and class_weight = 'balanced'; C was tuned.

2.5.4 Decision Tree algorithm

The DT recursively partitions the feature space into homogeneous regions using decision rules based on the Gini impurity criterion [6]:

$$Gini(t) = 1 - \sum_{k=1}^{K} p(k|t)^2 \quad (4)$$

where, $p(k|t)$ is the proportion of class k samples at node t, and K is the number of classes. DT was configured with class_weight = 'balanced', criterion = 'gini' and random_state = 42; to control overfitting, max_depth was tuned. Figure 3 illustrates the structure of a DT (Figure 3(a)) and, for comparison, the sequential learning scheme of GB described in Section 2.5.2 (Figure 3(b)).

2.6 Hyperparameter configuration

Hyperparameters were tuned by nested cross-validation. Within each outer training partition, an inner GridSearchCV selected the configuration that maximized the F1-Score (classification) or R^2 (regression), and the selected pipeline was then evaluated on the corresponding outer test partition. The inner splitting scheme always matched the outer one: shuffled stratified folds in the stratified arm, and forward-chaining TimeSeriesSplit folds in the chronological arm. Hyperparameters used in the chronological arm were therefore never chosen under a random split. Table 4 reports the fixed settings of each model, and Table 5 reports the values evaluated for each tuned hyperparameter and the value

selected in every outer fold of both strategies.

One hyperparameter was tuned for each model: max_depth for RF (None or 10), GB (3 or 5, with learning_rate fixed at 0.1) and DT (10 or None), and C for SVM (1 or 10). The selected values varied little between folds (Table 5). The clearest differences between strategies concerned GB in regression (max_depth = 5 in every stratified fold and 3 in every chronological fold) and SVM in classification (C = 10 in every stratified fold and C = 1 in four of the five chronological folds).

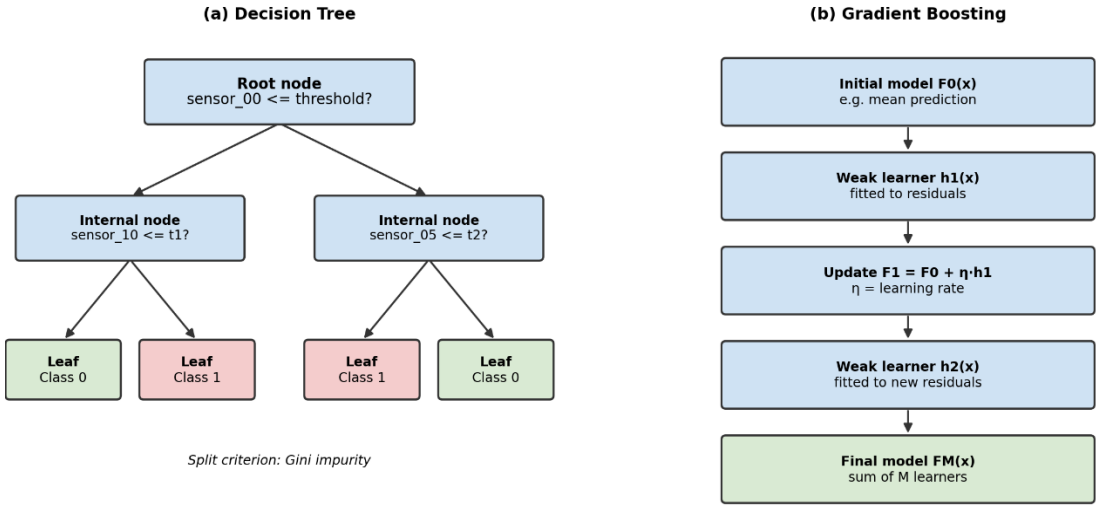


Figure 3. Decision Tree (DT) structure (left) and Gradient Boosting (GB) sequential learning (right)

Table 4. Fixed hyperparameter settings of each model

Parameter	RF	GB	SVM	DT
n_estimators	100	100	N/A	N/A
max_depth	Tuned	Tuned	N/A	Tuned
learning_rate	N/A	0.1	N/A	N/A
kernel	N/A	N/A	RBF	N/A
C (regularization)	N/A	N/A	Tuned	N/A
gamma	N/A	N/A	scale	N/A
max_features	sqrt	sqrt	N/A	N/A
class_weight	balanced	N/A	balanced	balanced
criterion	gini	friedman_mse	N/A	gini
random_state	42	42	42	42

Note: Settings of the classification models; 'Tuned' denotes a hyperparameter selected by nested cross-validation (values in Table 5). Regression used the corresponding scikit-learn regressors with the same settings, except that class_weight does not apply, RF and DT used the default squared-error criterion, and GB used the default max_features. Gradient Boosting (GB), Random Forest (RF).

Table 5. Tuned hyperparameters: values evaluated and value selected in each outer fold

Model	Task	Tuned Hyperparameter	Values Evaluated	Selected (Stratified Folds)	Selected (Chronological Folds)
RF	Classification	max_depth	None, 10	None (4), 10 (1)	None (4), 10 (1)
RF	Regression	max_depth	None, 10	None (5)	10 (3), None (2)
GB	Classification	max_depth	3, 5	5 (3), 3 (2)	3 (3), 5 (2)
GB	Regression	max_depth	3, 5	5 (5)	3 (5)
SVM	Classification	C	1, 10	10 (5)	1 (4), 10 (1)
SVM	Regression	C	1, 10	10 (5)	10 (5)
DT	Classification	max_depth	10, None	10 (5)	10 (5)
DT	Regression	max_depth	10, None	10 (4), None (1)	10 (5)

Note: Values selected by an inner GridSearchCV with three inner folds of the same type as the outer scheme (StratifiedKFold with shuffle = True and random_state = 42, or TimeSeriesSplit), using the F1-Score for classification and R^2 for regression. Numbers in parentheses are the number of the five outer folds in which each value was selected. n_estimators = 100 (RF and GB) and learning_rate = 0.1 (GB) were fixed. Decision Tree (DT), Gradient Boosting (GB), Random Forest (RF), Support Vector Machine (SVM).

2.7 Performance metrics

2.7.1 Classification metrics

Classification models were evaluated using the following four metrics [22]:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (5)$$

$$Precision = TP/(TP + FP) \quad (6)$$

$$Recall = TP/(TP + FN) \quad (7)$$

$$F1\text{-}Score = 2 \times (Precision \times Recall)/(Precision + Recall) \quad (8)$$

where, TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives. Results are reported as mean ± standard deviation over the five outer folds of each validation strategy [13].

2.7.2 Regression metrics

The following four metrics were used for the regression models [23]:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (9)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (10)$$

$$RMSE = \sqrt{[\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2]} \quad (11)$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (12)$$

where, y_i are actual values, $\hat{y}_i$ are predicted values, $\bar{y}$ is the mean of actual values, and n is the number of observations. Because R^2 can be misleading when the target varies little within a test window, mean absolute error (MAE) and root mean squared error (RMSE) are reported alongside it.

2.8 Computational environment and procedure

The revised analysis was implemented in Python 3 with scikit-learn 1.8.0, pandas 3.0.2, NumPy 2.4.4, SciPy 1.17.1 and Matplotlib 3.10.8. Class weighting was set to "balanced" for RF, SVM and DT, while GB was trained without class weighting, as reported in Table 4. Reproducibility was ensured by fixing random_state = 42 across all stochastic components, and the complete analysis script is available from the authors on request.

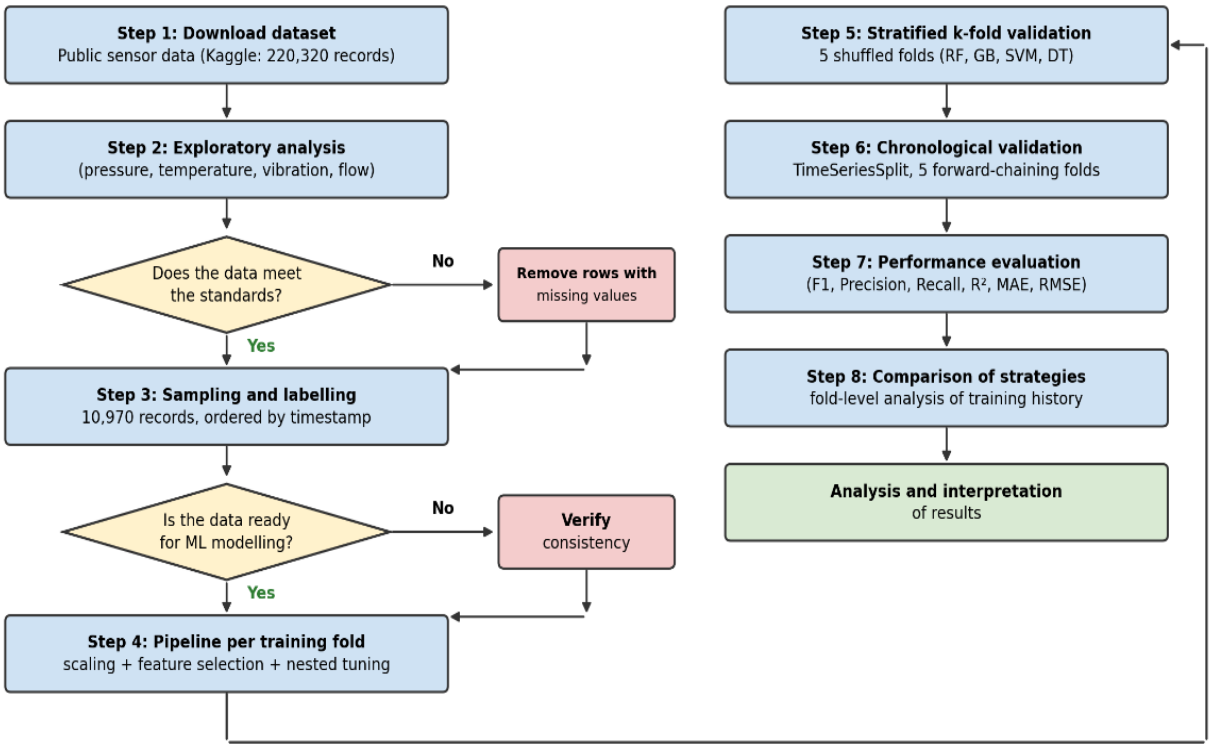


Figure 4. Research procedure flowchart

Figure 4 details the steps that were followed during the research process with the objective of evaluating machine learning models for the classification and prediction of the SPI in underground mining auxiliary services.

3. RESULTS

Following the methodology and objectives of the research, this section presents the performance of the RF, GB, SVM, and DT models under both classification and regression approaches, accompanied by a specific discussion of each result.

3.1 Exploratory analysis and class definition

Using the original machine_status labels, the sample contains 8,000 records of Class 0 (Optimal) and 2,970 of Class 1 (Degraded), a proportion of 72.9% to 27.1%. Figure 5 shows the scatter plots between the most relevant sensor variables, revealing clearly differentiated groupings between classes.

As shown in Figure 5, Class 1 records (red) tend to cluster in regions with extreme suction pressure and temperature values, while Class 0 records (blue) display a more compact distribution within normal operating ranges. This visual separation confirms that the selected sensor variables contain sufficient discriminative information to classify the operational state of the pumps. The distinct clustering pattern observed-particularly between suction pressure and fluid temperature-validates the choice of these variables as primary predictors, as degraded pumps consistently exhibit abnormal pressure-temperature relationships compared to optimally operating units. The moderate class imbalance (73%/27%) justifies the use of class_weight='balanced' in RF, SVM, and DT, and the selection of F1-Score as the primary evaluation metric.

The box plots in Figure 6 reinforce this finding, showing the distribution of each variable by operational class.

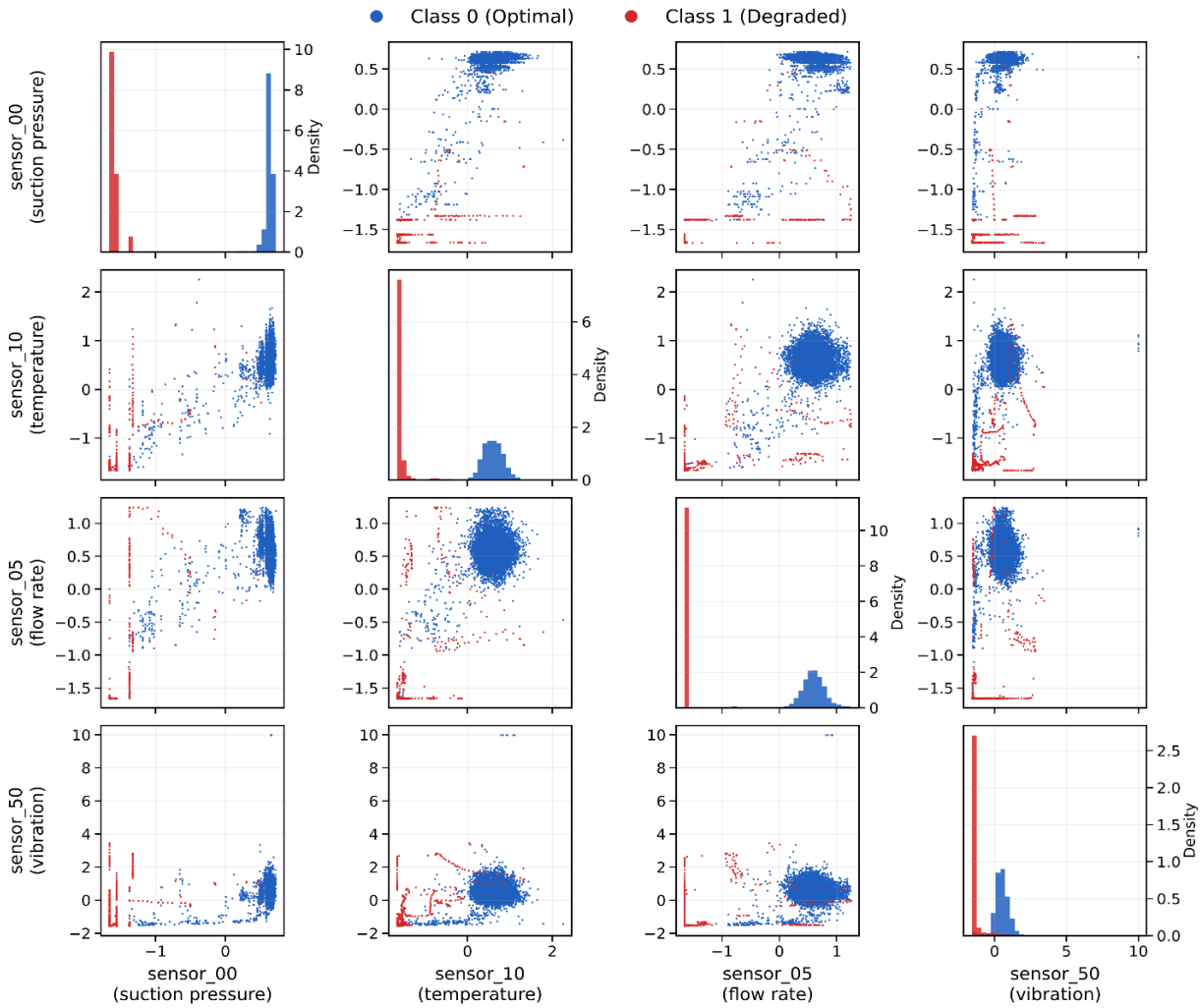


Figure 5. Scatter plot by class-operational sensor variables

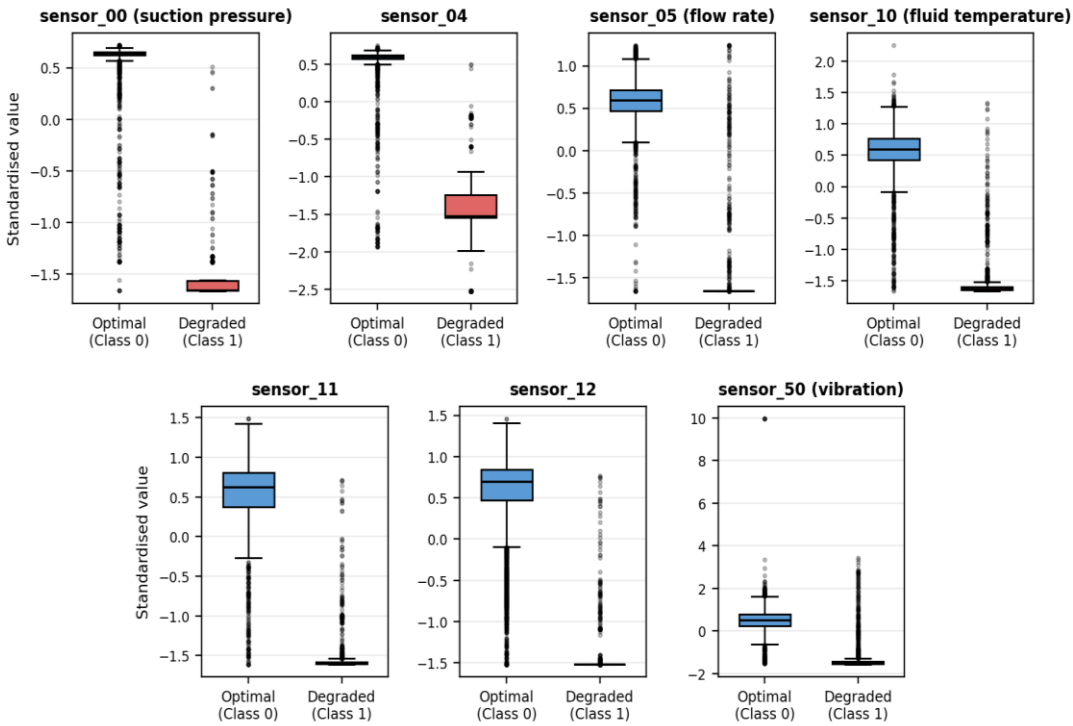


Figure 6. Distribution by class of the seven variables selected for classification in the stratified folds (standardized values)

Figure 6 shows, by class, the distribution of the seven variables selected for classification in the stratified folds. All seven separate the classes strongly: for none of them do more than 1% of degraded records fall within the interquartile range of the optimal class, and the standardized mean difference ranges from 3.2 for vibration (sensor_50) to 11.2 for suction pressure (sensor_00). Suction pressure, sensor_11 and fluid temperature (sensor_10) provide the clearest separation, whereas vibration overlaps most and is better regarded as a complementary signal. This strong separability explains why classification is straightforward once degradation episodes are represented in the training data, and why the stratified scores in Table 6 are so high.

Table 6. Performance metrics-classification (stratified k-fold, five outer folds)

Model	Accuracy	Precision	Recall	F1-Score
RF	0.9979 ± 0.0012	0.9950 ± 0.0041	0.9973 ± 0.0008	0.9961 ± 0.0023
GB	0.9982 ± 0.0010	0.9953 ± 0.0040	0.9980 ± 0.0007	0.9966 ± 0.0019
SVM	0.9964 ± 0.0008	0.9887 ± 0.0016	0.9983 ± 0.0015	0.9935 ± 0.0014
DT	0.9969 ± 0.0017	0.9936 ± 0.0033	0.9949 ± 0.0035	0.9943 ± 0.0031

Note: Values are mean ± standard deviation over five shuffled outer folds, with scaling, feature selection and hyperparameter tuning nested inside each training fold. Random Forest (RF); Gradient Boosting (GB); Support Vector Machine (SVM); Decision Tree (DT).

3.2 Training and validation of classification models

The four models were first evaluated with stratified k-fold cross-validation (five shuffled outer folds), with scaling, feature selection and hyperparameter tuning nested inside each training fold. Table 6 reports the results. As Section 3.6 shows, these stratified estimates are optimistic because random shuffling ignores the temporal order of the observations; they are reported as the reference against which the chronological evaluation is compared, not as estimates of prospective performance.

Under stratified k-fold all four models reached very high scores, with F1-Scores between 0.9935 (SVM) and 0.9966 (GB) and accuracies between 0.9964 and 0.9982. The differences among models are of the same order as the fold-to-fold standard deviations (0.0014 to 0.0031), so no model can be regarded as superior. SVM showed the highest recall (0.9983) and the lowest precision (0.9887), consistent with a more conservative decision boundary that flags ambiguous records as degraded.

3.3 Confusion matrices and decision boundaries

The confusion matrices in Figure 7 were generated by pooling the out-of-fold predictions of the five stratified outer folds, so that every one of the 10,970 records contributes exactly once, as a prediction made by a model that did not see it during training.

Figure 7 shows that misclassifications are rare under stratified validation for all four models. GB produced the fewest errors overall (14 false positives and 6 false negatives), followed by RF (15 and 8). SVM produced the fewest false negatives (5) but the most false positives (34), consistent with its more conservative boundary, whereas DT produced the most false negatives (15). These stratified results describe the ability to recognize degradation episodes already represented in training; Section 3.6 shows that they do not carry over to periods for which no prior degradation history is available.

The decision boundaries in Figure 8, visualized through PCA projection to two principal components, illustrate the spatial strategy of each model.

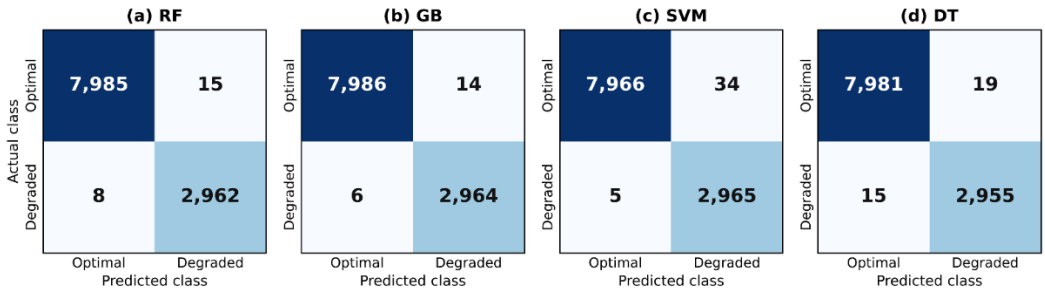


Figure 7. Confusion matrices from pooled out-of-fold predictions under stratified validation: (a) Random Forest (RF); (b) Gradient Boosting (GB); (c) Support Vector Machine (SVM); (d) Decision Tree (DT)

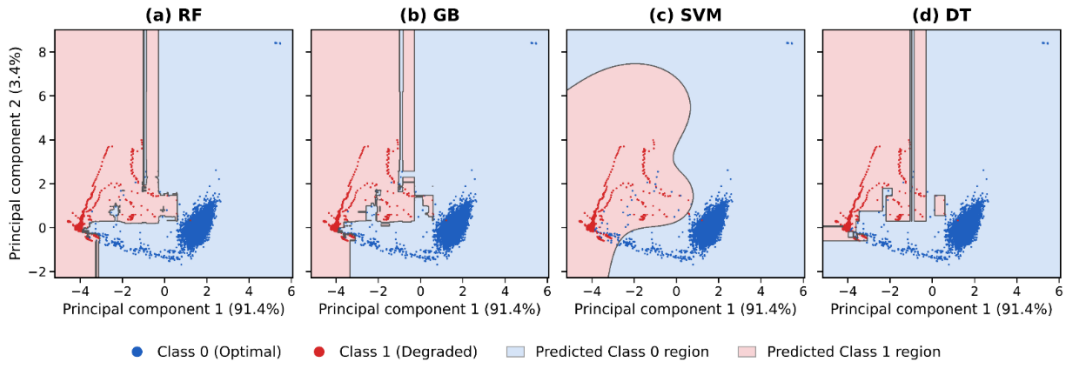


Figure 8. Decision boundaries-PCA 2D projection: (a) Random Forest (RF); (b) Gradient Boosting (GB); (c) Support Vector Machine (SVM); (d) Decision Tree (DT)

Figure 8 illustrates how each model partitions a two-dimensional PCA projection of the selected variables. RF (Figure 8(a)), GB (Figure 8(b)) and DT (Figure 8(d)) produce irregular, axis-aligned partitions, whereas SVM (Figure 8(c)) produces a smooth curved boundary, consistent with its RBF kernel and margin-maximization objective. Because the projection retains only two components of a seven-dimensional space, these maps are qualitative illustrations of model behavior rather than evidence of relative performance.

3.4 Training and validation of regression models

For the regression approach, the target variable was the SPI, expressed as a percentage (%), derived from the min-max normalization of suction pressure (sensor_00) on a 0-100% scale. To prevent information leakage, suction pressure (sensor_00)-the variable from which this target is derived-was excluded from the regression predictors, so that the SPI is estimated from the seven variables selected in each training fold among the remaining 50 sensors. It should be noted explicitly that the SPI, as defined here, is a single-sensor-derived index reflecting suction-pressure behavior; it is not a composite or independently validated measure of overall pump efficiency, hydraulic power, or remaining useful life, and the regression results below should be interpreted within that scope. The four models were evaluated using 5-fold cross-validation. Full precision results are presented in Table 7.

Under shuffled k-fold, the four models reconstructed the SPI from the remaining sensors with R^2 between 0.9959 (SVM) and 0.9977 (RF), MAE between 1.10 and 1.33 and RMSE between 2.02 and 2.66 SPI units. As in classification, the differences among models are small relative to their fold-to-fold variability, and the chronological evaluation in Section 3.6 shows that these values overstate prospective performance.

Figure 9 compares predicted and actual SPI values using the pooled out-of-fold predictions of the stratified arm. For all four models the points lie close to the diagonal, and most of the error is concentrated in the intermediate range of the index: for the 207 records with an SPI between 20% and 80%, RMSE ranged from 6.80 (RF) to 10.51 (SVM), against 1.82 to 2.27 outside that range. This intermediate range corresponds to the transition between optimal and degraded operation, where early detection would be most valuable, and it is also where the models are least accurate.

Table 7. Performance metrics-regression of the Suction Pressure Index (SPI) (shuffled k-fold, five outer folds)

Model	R^2	MAE	MSE	RMSE
RF	0.9977 ± 0.0003	1.0958 ± 0.0299	4.1064 ± 0.4455	2.0234 ± 0.1110
GB	0.9974 ± 0.0006	1.1859 ± 0.0318	4.6150 ± 0.9571	2.1360 ± 0.2292
SVM	0.9959 ± 0.0008	1.3266 ± 0.0449	7.1301 ± 1.2412	2.6604 ± 0.2286
DT	0.9969 ± 0.0003	1.1858 ± 0.0414	5.5516 ± 0.7300	2.3512 ± 0.1534

Note: R^2: coefficient of determination; MAE: mean absolute error; MSE: mean squared error; RMSE: root mean squared error. MAE and RMSE are expressed in percentage points (pp), while MSE is expressed in squared percentage points (pp^2). Values are reported as mean ± standard deviation over five shuffled outer folds. Function pressure (sensor_00) was excluded from the predictors because the SPI is derived from it. RF: Random Forest; GB: Gradient Boosting; SVM: Support Vector Machine; DT: Decision Tree.

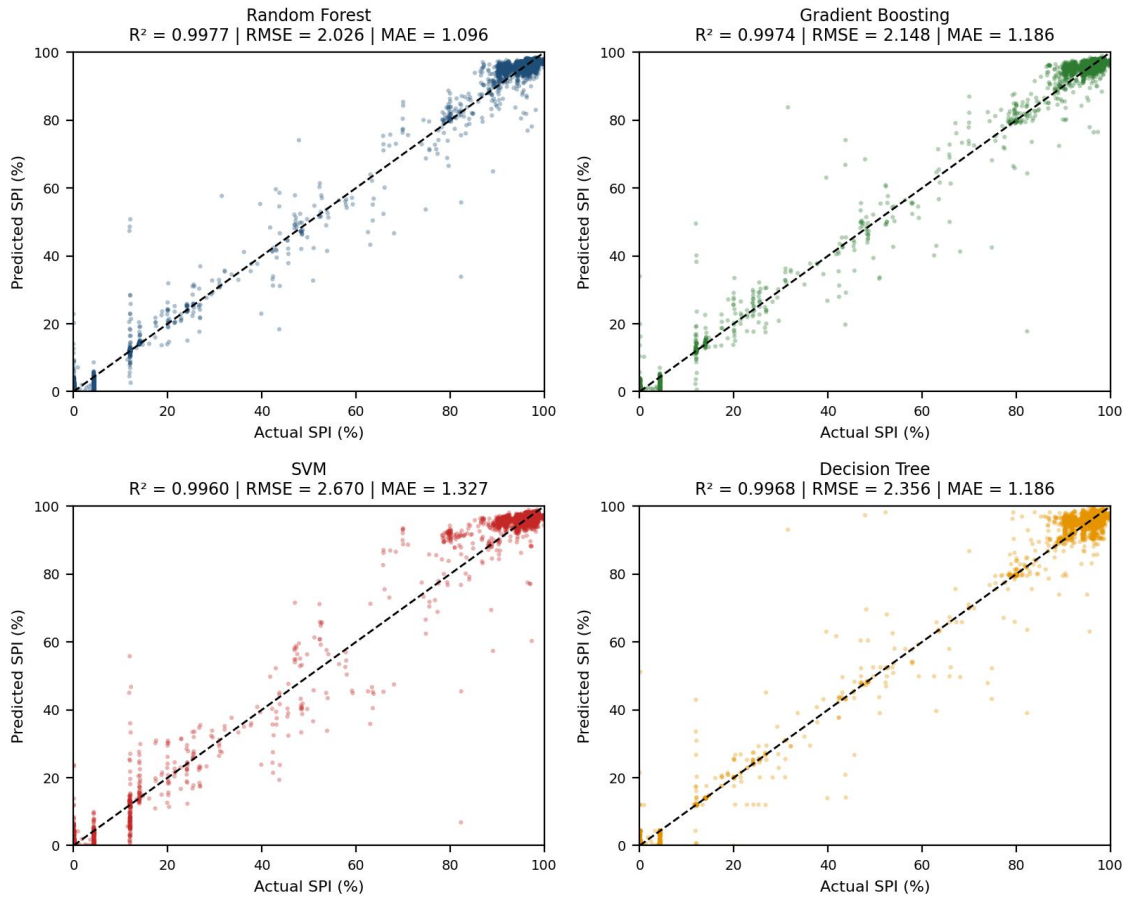


Figure 9. Predicted vs. actual Suction Pressure Index (SPI) from pooled out-of-fold predictions under stratified validation: (a) Random Forest (RF); (b) Gradient Boosting (GB); (c) Support Vector Machine (SVM); (d) Decision Tree (DT)

3.5 Comparative analysis

Figure 10 presents a graphical comparison of the stratified-arm metrics of the four models for both the classification and the regression tasks.

Figure 10 summarizes the stratified-arm metrics of Tables 6 and 7. The vertical axis of Figure 10(a) starts at zero, so the bars convey correctly that the four models are practically indistinguishable under this validation strategy. Figure 10(b) shows the corresponding regression errors, with MAE between 1.10 and 1.33 and RMSE between 2.02 and 2.66 SPI units.

3.6 Effect of validation strategy on reported performance

To examine whether the stratified estimates reflect prospective predictive capability, the same 10,970-record sample was re-evaluated with chronological forward-chaining validation (TimeSeriesSplit, five folds), in which every test fold lies later in time than its training data. The data, candidate features, pipelines and hyperparameter grids were identical in both arms; only the splitting scheme differed, and it was applied consistently to the inner tuning loop as well. Table 8 compares the two strategies.

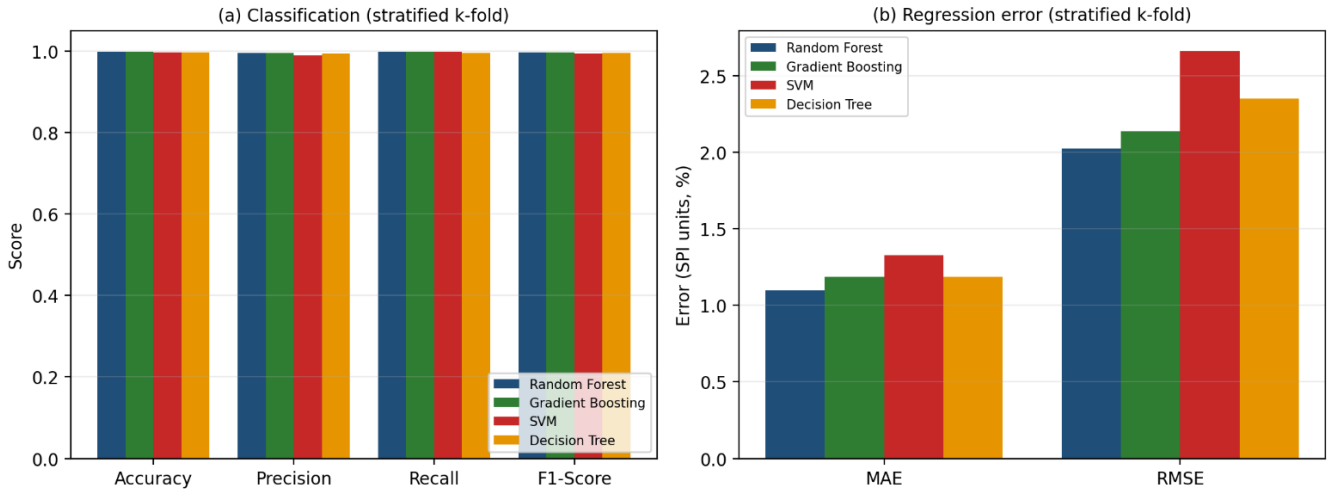


Figure 10. Performance under stratified validation: (a) classification metrics; (b) regression errors (mean absolute error (MAE) and root mean squared error (RMSE))

Table 8. Performance of each model under stratified and chronological validation on the same 10,970-record sample

Model	F1 Stratified	F1 Chronological (4 Evaluable Folds)	F1 Fold 2 (37 Deg. in Training)	F1 Folds 3-5 (781-1,596 Deg.)	RMSE Stratified / Chronological
RF	0.9961	0.7332 ± 0.4223	0.003	0.9767	2.02 / 6.08
GB	0.9966	0.7358 ± 0.4236	0.003	0.9801	2.14 / 6.22
SVM	0.9935	0.7374 ± 0.4261	0.000	0.9832	2.66 / 12.23
DT	0.9943	0.7337 ± 0.4228	0.003	0.9773	2.35 / 4.84

Note: Stratified values are means over five shuffled folds. Chronological F1 is the mean ± SD over the four evaluable folds; fold 1 contains no degraded record in its test window, so F1 is undefined there. Fold 2 was trained with 37 degraded records and folds 3–5 with 781 to 1,596. Root mean squared error (RMSE) is expressed in Suction Pressure Index (SPI) units (%). Random Forest (RF); Gradient Boosting (GB); Support Vector Machine (SVM); Decision Tree (DT).

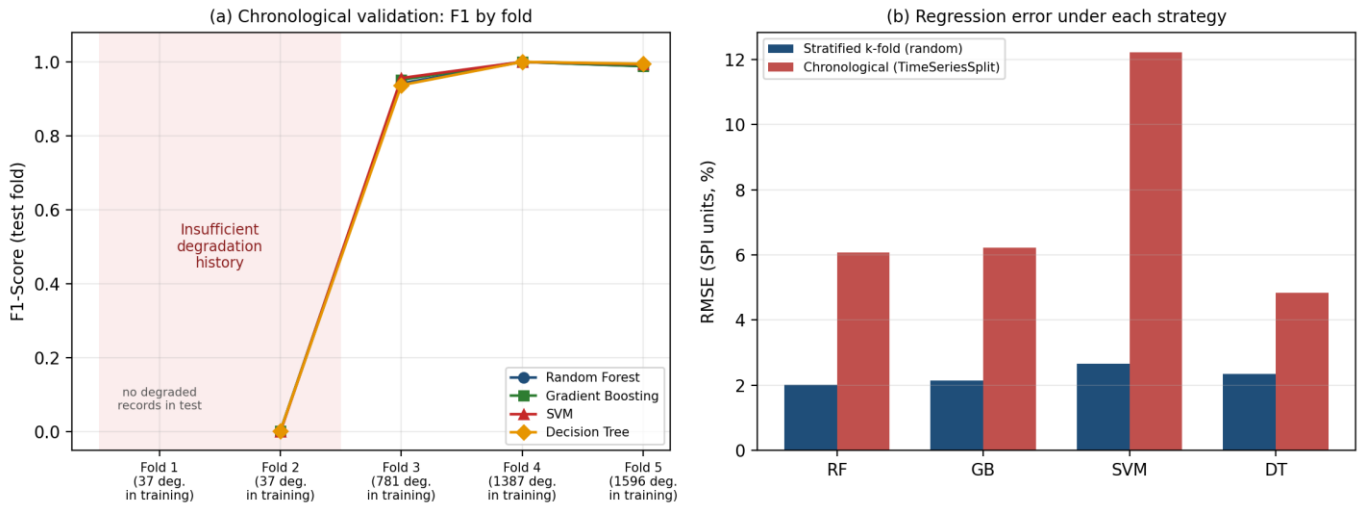


Figure 11. (a) F1-Score of each model in each chronological fold, with the number of degraded records available for training; (b) regression root mean squared error (RMSE) under stratified and chronological validation

Under chronological validation the mean F1-Score of every model fell to about 0.73, with standard deviations above 0.42, but this average conceals a sharply bimodal pattern (Figure 11(a)). In fold 1 the test window contains no degraded record, so F1 is undefined. In fold 2 the models were trained on a window containing only 37 degraded records and failed almost completely on a test window containing 744 (F1 no higher than 0.003; accuracy about 0.59, the proportion of optimal records), which indicates that they labeled virtually every test record as optimal. From fold 3 onwards, with 781 to 1,596 degraded records available for training, F1 ranged from 0.937 to 1.000 for all four models. The chronological evaluation therefore shows that the models recognize degradation in later periods only after comparable degradation episodes have been observed during training.

This behavior follows from the temporal structure of the data. Degradation records are concentrated in a small number of episodes rather than distributed uniformly over the monitoring period: the first test window contains none, whereas later windows contain between 209 and 1,374. Under random shuffling, minute-level records from every episode are spread across training and test partitions, so each test fold contains observations nearly identical to training observations from the same episode. The stratified estimates thus measure the ability to recognize episodes already represented in training, and they conceal the cold-start behavior exposed by chronological validation. The regression task shows the same dependence: RMSE rises from 2.02–2.66 SPI units under stratified validation to 4.84–12.23 under chronological validation (Figure 11(b)). The coefficient of determination is uninformative in fold 1, where the pump operated normally throughout and the SPI varied very little (standard deviation 1.18, against 30.95-46.52 in the other test windows): R^2 values between -6.59 and -5.06 there correspond to absolute errors of only about 3 SPI units (RMSE 2.90-3.24). Median R^2 across the chronological folds ranged from 0.86 (SVM) to 0.98 (GB).

Because the four algorithms share the same data and splits, they cannot be treated as independent replicates, and no statistical test across models is reported. Within each model, the comparison between strategies rests on the fold-level results in Table 8 and Figure 11; with five outer folds per arm and a bimodal chronological distribution, formal significance tests would have very low power, so the results are reported descriptively. The four algorithms behaved almost identically under both strategies, including the failure in fold 2 and the recovery from fold 3 onwards, which indicates that the dominant factor is the training history rather than the choice of algorithm.

4. DISCUSSION

4.1 Discussion of results

The central result of this study is that the validation strategy, rather than the choice of algorithm, determines the performance reported for pump condition monitoring on this dataset. Under stratified cross-validation all four models appear nearly perfect and practically indistinguishable; under chronological validation all four fail in the same period and recover in the same period. Published studies reporting near-perfect metrics for random splits of time-series sensor data, including those that identify ensemble methods as superior [7, 8, 11, 12], should therefore be read with the validation design in mind.

This finding does not contradict the physical relevance of the selected variables discussed below; it shows that the information they carry can be exploited only once a model has observed how those variables evolve during degradation.

Under stratified validation SVM showed the highest recall (0.9983) at the cost of the lowest precision (0.9887). In underground mining, where an undetected pump failure can flood workings, such conservative behavior would be preferable. However, under chronological validation SVM failed in fold 2 exactly as the other models did (F1 = 0.000), so this advantage cannot be relied upon when the training history lacks degradation episodes [10].

In regression, the stratified R^2 values (RF 0.9977, GB 0.9974, DT 0.9969, SVM 0.9959) are again practically indistinguishable. Under chronological validation the tree-based models retained median R^2 between 0.976 and 0.982 and RMSE between 4.84 and 6.22 SPI units, whereas SVM degraded more (median R^2 0.856, RMSE 12.23), mainly because of a large error in fold 2 (RMSE 34.52).

The SPI is itself a transformation of suction pressure, which was excluded from the regression predictors to avoid circularity; the regression results therefore describe how well the remaining sensors reconstruct suction-pressure behavior, and are subject to the scope limitation stated in Section 3.4.

Beyond comparative performance, the selected variables admit a direct physical interpretation that links the model outputs to maintenance decisions. Suction pressure (sensor_00) was the most discriminative variable because a sustained pressure drop is the primary hydraulic signature of impeller wear, seal leakage, or partial clogging, all of which reduce a pump's capacity to lift water. Fluid temperature (sensor_10) was also strongly discriminative because abnormal heating reflects increased mechanical friction from bearing degradation or inadequate lubrication, and often precedes catastrophic failure. Flow rate captures the combined hydraulic effect of these deteriorations, decreasing as the pump loses efficiency. Vibration (sensor_50), although more overlapping between classes, provides complementary early-warning information, since rising vibration amplitude is a well-established indicator of cavitation, misalignment, and imbalance. This physical reading shows how a dominant pressure–temperature signal and a noisier vibration signal complement each other, and it allows maintenance teams to translate a model alert into a specific inspection target-for example, prioritizing seal and impeller inspection when a pressure-driven alert is raised, versus bearing and lubrication checks when temperature and vibration dominate.

4.2 Practical implications and implementation considerations

The results suggest how such models might be deployed in underground mining, but they do not constitute a validated application: no data from an operating mine were available, and the chronological evaluation showed that performance depends on the degradation history available for training. Any deployment should therefore be treated as a pilot, and several conditions would have to be met: IoT sensor infrastructure for continuous data acquisition, an initial period in which degradation episodes are recorded and labeled before the model is relied upon, periodic retraining to incorporate new episodes and compensate for sensor drift, and training of maintenance personnel to interpret the outputs.

To make the deployment pathway concrete, a trained model could be integrated into an intelligent maintenance system organized as a five-stage workflow. (1) Data acquisition: IoT sensors installed on each drainage pump continuously stream suction pressure, temperature, flow rate, and vibration at regular intervals to an edge gateway. (2) Preprocessing: incoming records are cleaned of missing and duplicate values and standardized with the same StandardScaler parameters fitted during training, ensuring consistency between deployment and model development. (3) Model prediction: the trained model classifies each record as Optimal or Degraded and, in parallel, estimates a continuous SPI value. (4) Condition evaluation: predictions are aggregated over a sliding time window to filter transient noise and confirm a genuine degradation trend rather than an isolated reading. (5) Maintenance decision support: when the confirmed condition crosses a preset risk threshold, the system issues a prioritized alert to the maintenance team, recommending inspection of the components implicated by the dominant sensor signal. Periodic retraining with newly labeled site data compensates for sensor drift and changing operating conditions, closing the loop between monitoring and action.

5. CONCLUSIONS

The training and evaluation of four machine learning models (RF, GB, SVM, and DT) for the pump condition classification and SPI estimation in underground mining auxiliary services yielded the following key findings.

Under randomly shuffled stratified cross-validation, all four models (RF, GB, SVM and DT) appeared almost perfect, with F1-Scores between 0.9935 and 0.9966 and regression R^2 between 0.9959 and 0.9977, and no model was clearly superior. When the same 10,970-record sample was evaluated chronologically, with feature selection and hyperparameter tuning nested inside each training window, performance depended on the degradation history available for training: every model failed when only 37 degraded records had been observed (F1 no higher than 0.003) and performed well once 781 or more had been observed (F1 between 0.937 and 1.000).

The main contribution of this study is methodological. It shows that random cross-validation of minute-level pump sensor data conceals a cold-start dependence: the models recognize degradation only after comparable episodes have appeared in their training history. Reporting accuracy alone would obscure this, since in the failed fold accuracy remained near 0.59, simply reflecting the proportion of optimal records. For predictive maintenance in mining, where an undetected failure can flood workings and endanger personnel, we recommend that models be validated chronologically, that performance be reported per time window together with the number of prior degradation episodes, and that no model be relied upon until its training history contains representative failure episodes.

5.1 Limitations and future research directions

This study is based on a public industrial pump dataset, which may limit direct generalization to specific underground mining contexts with different geological and hydrological conditions.

Furthermore, although the physical degradation mechanisms are expected to be analogous, underground mining introduces operating conditions not fully represented in the industrial dataset. Mine water is frequently acidic and carries abrasive suspended solids that accelerate impeller and seal wear; ambient humidity, elevated temperature, and confined ventilation alter the thermal baseline of the equipment; and drainage demand fluctuates with groundwater inflow, imposing variable and intermittent operating loads. These mining-specific factors may shift the absolute sensor values and the rate of degradation, so the thresholds and model parameters obtained here should be recalibrated against site-specific data before operational deployment.

A further limitation concerns the amount of degradation history. The dataset contains a limited number of degradation episodes concentrated in a few periods, so the chronological evaluation rests on five folds, one of which has no degraded record in its test window and one of which is trained with very few. The fold-level results are therefore reported descriptively, and the recovery observed from fold 3 onwards should be confirmed on longer records containing more independent episodes.

The chronological forward-chaining validation reported in Section 3.6 addresses temporal leakage within this dataset, but two further forms of validation were not possible: splits grouped by operating episode, which would require episode labels that the dataset does not provide, and validation against an entirely independent dataset, since no public sensor dataset for mine dewatering pumps is currently available. External validation on data from operating mines is the most important direction for subsequent work.

The generalization claim toward underground mining also remains untested. The physical analogy between industrial centrifugal pumps and mining drainage pumps is argued on mechanistic grounds, but no measurements from an operating mine were available to verify that the degradation signatures transfer in practice.

Future research should validate this methodology using real-time sensor data from actual underground mining operations in Peru, expand the feature set to include differential pressure and multi-axis vibration, and integrate the validated model into automated real-time monitoring platforms.

ACKNOWLEDGMENT

The authors would like to thank Universidad Privada del Norte for institutional support and Ing. Barrera Gutierrez, Ramon Enrique, for his valuable academic guidance and technical recommendations throughout this research.

REFERENCES

[1] Sanmiquel, L., Rossell, J.M., Vintro, C. (2015). Study of Spanish mining accidents using data mining techniques. Safety Science, 75: 49-55. https://doi.org/10.1016/j.ssci.2015.01.016

[2] Werbinska-Wojciechowska, S., Rogowski, R. (2025). Proactive maintenance of pump systems operating in the mining industry-a systematic review. Sensors, 25(8): 2365. https://doi.org/10.3390/s25082365

[3] Noriega-Vidal, E.M., Narvaez-Valdivia, J.W., Huancas-Morey, M.A., Hernandez-Puyo, D.A., Effio-Quezada, W. (2026). Estimation of decision boundaries for critical

zone classification in a polymetallic tailings dam using machine learning. International Journal of Computational Methods and Experimental Measurements, 14(1): 156-171. https://doi.org/10.56578/ijcmem140110

[4] Alghtus, K.M.A., Gannan, A., Alhajri, K. M., Al Jubouri, A.L.A., Al-Janahi, H.A.I. (2025). Short-horizon predictive maintenance of industrial pumps using time-series features and machine learning. arXiv preprint, arXiv:2508.19974. https://doi.org/10.48550/arXiv.2508.19974

[5] Silva, J.P.S., Maitelli, A.L. (2025). K-nearest neighbors for anomaly detection and predictive maintenance in water pumping systems. Sensors, 25(11): 3532. https://doi.org/10.3390/s25113532

[6] Aliyu, R., Mokhtar, A.A., Hussin, H. (2024). Classification of pump failure using a decision tree technique. In ICREEM 2022. Springer, Singapore, pp. 319-336. https://doi.org/10.1007/978-981-99-5946-4_26

[7] Mallioris, P., Diamantis, E., Bialas, C., Bechtsis, D. (2024). Predictive maintenance framework for assessing health state of centrifugal pumps. IAES International Journal of Artificial Intelligence, 13(1), 850-862. https://doi.org/10.11591/ijai.v13.i1.pp850-862

[8] Chang, K., Park, S.H. (2024). Random forest-based multi-faults classification modeling and analysis for intelligent centrifugal pump system. Journal of Mechanical Science and Technology, 38: 11-20. https://doi.org/10.1007/s12206-023-1202-2

[9] Arukhe, J.O., Kalu-Ulu, T.C., Anazi, A.F., Almarri, H.M. (2024). Predicting the lifespan of electric submersible pumps using random-forest machine-learning. In SPE Asia Pacific Oil and Gas Conference and Exhibition, p. D021S012R007. https://doi.org/10.2118/221203-MS

[10] Chambi, N., Sanga, C., Ortiz, J., et al. (2025). Predictive maintenance in underground mining equipment using artificial intelligence. Eng, 6(10): 261. https://doi.org/10.3390/eng6100261

[11] Yang, Y.Q., Wang, H.J. (2025). Random forest-based machine failure prediction: A performance comparison. Applied Sciences, 15(16): 8841. https://doi.org/10.3390/app15168841

[12] Aly, M., Behiry, M.H. (2025). Enhancing anomaly detection in IoT-driven factories using logistic boosting, random forest, and SVM: A comparative machine learning approach. Scientific Reports, 15: 23694. https://doi.org/10.1038/s41598-025-08436-x

[13] Szeghalmy, S., Fazekas, A. (2023). A comparative study of the use of stratified cross-validation and distribution-balanced stratified cross-validation in imbalanced learning. Sensors, 23(4): 2333. https://doi.org/10.3390/s23042333

[14] Lumumba, V.W., Kiprotich, D., Mpaine, M.L., Makena, N.G., Kavita, M.D. (2024). Comparative analysis of cross-validation techniques: LOOCV, K-folds, and repeated K-folds in machine learning models. SSRN, p. 11. https://doi.org/10.2139/ssrn.5266507

[15] Bhagat, M., Bakariya, B. (2025). A comprehensive review of cross-validation techniques in machine learning. International Journal of Science and Technology (IJSAT), 16(1): 1305. https://doi.org/10.71097/IJSAT.v16.i1.1305

[16] UNKNOWNCLASS. (2018). Pump Sensor Data. Kaggle. https://www.kaggle.com/datasets/nphantawee/pump-sensor-data

[17] Shao, Z.X., Kumral, M. (2024). Implementing Gaussian process modelling in predictive maintenance of mining machineries. Mining Technology: Transactions of the Institutions of Mining and Metallurgy, 133(4): 348-368. https://doi.org/10.1177/25726668241275434

[18] Geron, A. (2023). Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow. O'Reilly Media, Sebastopol.

[19] Alaminos-Fernandez, A.F. (2022). Árboles de Decisión en R con Random Forest. Limencop, Alicante, Spain.

[20] Khan, A.A., Chaudhari, O., Chandra, R. (2024). A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. Expert Systems with Applications, 244: 122778. https://doi.org/10.1016/j.eswa.2023.122778

[21] Betancourt, G.A. (2005). Las maquinas de soporte vectorial (SVMs). Scientia et Technica, 1(27): 67-72.

[22] Powers, D.M.W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. Journal of Machine Learning Technologies, 2(1): 37-63. https://bioinfopublication.org/files/articles/2_1_1_JMLT.pdf.

[23] Chicco, D., Warrens, M.J., Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, 7: e623. https://doi.org/10.7717/peerj-cs.623