Engineering Frontiers
Vol. 1, No. 2, 2025 42
Research Article

Machine Learning-Based Fault Detection for UR3 Collaborative Robot: A Multimodal Data Analysis Approach

Shida Liu [a] *
[a] College of Saint Petersburg Joint Engineering, Xuzhou University of Technology, Xuzhou 221018, China

Abstract
Collaborative robots (cobots) are increasingly deployed in industrial environments, necessitating robust fault detection systems to ensure operational safety and efficiency. This study presents a comprehensive machine learning framework for fault detection in UR3 collaborative robots using multimodal sensor data. We analyzed a dataset comprising 7,409 samples with 20 features, including joint currents, temperatures, and velocities. Five machine learning algorithms were evaluated: Logistic Regression, Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), and Decision Tree. Our results demonstrate that KNN achieved the best overall performance with an accuracy of 94.42%, F1-score of 0.446, and AUC of 0.812. Feature importance analysis revealed that joint currents (J3, J2) and joint velocities (J5) are the most critical indicators for fault prediction. Dimensionality reduction techniques (PCA and t-SNE) confirmed distinct separability between normal and fault conditions. This work provides valuable insights for developing predictive maintenance systems in collaborative robotics applications.
Keywords: Collaborative robot; Fault detection; Machine learning; Predictive maintenance; Multimodal sensor fusion; Feature importance

1 Introduction

Collaborative robots represent a transformative technology in modern manufacturing, enabling safe human-robot interaction without traditional safety barriers. Unlike conventional industrial robots, cobots are designed with inherent safety features, including force-torque sensors and protective stop mechanisms, to prevent injuries during physical contact with human operators(Faccio et al., 2023; Fernandez-Vega et al., 2025). However, the operational reliability of these systems remains a critical concern, as unexpected failures can lead to production downtime, economic losses, and potential safety hazards.

The UR3 robot, manufactured by Universal Robots, exemplifies the state-of-the-art in lightweight collaborative robotics. With a payload capacity of 3 kg and six degrees of freedom, it is widely deployed in precision assembly, material handling, and quality inspection tasks(Wijaya et al., 2024). Despite its sophisticated design, UR3 robots are subject to various failure modes, including mechanical wear, thermal degradation, sensor drift, and gripper malfunctions(Ali, 2025). Early detection of these anomalies is essential for implementing proactive maintenance strategies and minimizing unplanned downtime.

Traditional fault detection approaches in robotics rely on threshold-based monitoring of individual sensor signals(Wu et al., 2017). These methods, while simple to implement, often fail to capture complex fault patterns that manifest across multiple sensor modalities. Recent advances in machine learning have enabled data-driven approaches that can automatically learn discriminative features from high-dimensional sensor data(Sun & Ge, 2021). Several studies have demonstrated the efficacy of supervised learning algorithms for fault diagnosis in industrial equipment, but limited research has specifically addressed the unique challenges of

* Corresponding author. E-mail: liu27010599@163.com

https://doi.org/10.71411/ef.2025.v1i2.941

collaborative robot fault detection(Khalastchi & Kalech, 2018).

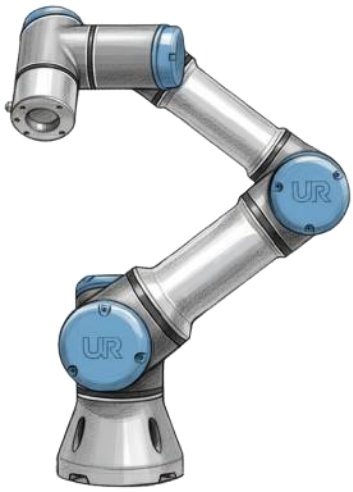


Figure 1. UR3 Robot Industrial Design Drawing.

The present study addresses this gap by developing a comprehensive machine learning framework for fault detection in UR3 robots. Our approach integrates multimodal sensor data, including joint currents, temperatures, and velocities, to construct a robust fault detection system. We systematically compare five mainstream machine learning algorithms and employ advanced feature engineering techniques to identify the most informative predictors of robot failures. Furthermore, we utilize dimensionality reduction methods to visualize fault patterns and gain mechanistic insights into failure modes.

The primary contributions of this work are threefold. First, we establish a benchmark evaluation of multiple machine learning algorithms for UR3 fault detection using real-world operational data. Second, we identify critical features and their relative importance through comprehensive feature importance analysis. Third, we demonstrate the practical applicability of our approach through detailed performance metrics and misclassification analysis. The findings provide actionable guidance for developing intelligent predictive maintenance systems in collaborative robotics applications.

2. Materials and Methods

2.1 Data Collection and Experimental Setup

The experimental dataset was collected from a UR3 collaborative robot deployed in a simulated industrial pick-and-place operation(Tyrovolas, 2024). Data acquisition was performed over an extended operational period, capturing both normal operating conditions and various fault scenarios. The dataset consists of 7,409 temporal samples, each characterized by 24 raw sensor measurements recorded at a sampling frequency appropriate for real-time monitoring.

The sensor suite includes current measurements from all six revolute joints (J0-J5), temperature readings from six thermal sensors (T0, J1-J5), velocity profiles for each joint, tool current, and operational cycle count. Additionally, two binary indicators were recorded: Robot_ProtectiveStop, which flags protective stop events triggered by force-torque thresholds, and grip_lost, which indicates gripper failure during object manipulation. These indicators serve as ground truth labels for fault detection, with a fault defined as any instance where either protective stop or gripper failure occurs.

The dataset exhibits significant class imbalance, with 6,891 normal samples (93.01%) and 518

https://doi.org/10.71411/ef.2025.v1i2.941

fault samples (6.99%), representing a realistic operational scenario where faults are relatively rare events. This imbalance ratio of approximately 13.3:1 necessitates careful consideration during model training and evaluation, as standard accuracy metrics may be misleading in such scenarios.

2.2 Data Preprocessing and Feature Extraction

Raw sensor data underwent a systematic preprocessing pipeline to ensure data quality and improve model performance. Missing values, which accounted for 1,018 instances across all features, were imputed using forward-fill and zero-fill strategies depending on the feature type. Infinite values were replaced with zeros to prevent numerical instabilities during subsequent processing. To address potential sensor drift and facilitate model convergence, all features were standardized using z-score normalization:

$$x_{\text{norm}} = \frac{x - \mu}{\sigma + \epsilon} \tag{1}$$

where $\boldsymbol{\mu}$and $\boldsymbol{\sigma}$represent the mean and standard deviation computed using omitnan options to handle any remaining missing values, and $\boldsymbol{\epsilon = 10^{-10}}$prevents division by zero.

From the 24 raw measurements, we extracted 20 engineered features for model training. The feature set comprises six joint currents ($\boldsymbol{I_{J0}}$ to $\boldsymbol{I_{J5}}$), six temperature measurements ($\boldsymbol{T_{T0}}$, $\boldsymbol{T_{J1}}$to $\boldsymbol{T_{J5}}$), six joint velocities ($\boldsymbol{\omega_{J0}}$ to $\boldsymbol{\omega_{J5}}$), tool current (I_{tool}), and operational cycle count. Temperature features are particularly informative as they reflect thermal accumulation from sustained mechanical stress and friction. Joint currents provide direct indicators of mechanical load and potential overload conditions. Velocity profiles capture dynamic characteristics of robot motion and can reveal degraded actuator performance.

2.3 Machine Learning Algorithms

Five supervised learning algorithms were implemented and evaluated for binary fault classification. The selection encompasses both linear and nonlinear methods to comprehensively assess model performance across different decision boundary complexities.

Logistic Regression serves as a baseline linear classifier, modeling the log-odds of fault occurrence as a linear combination of features. Despite its simplicity, logistic regression provides interpretable coefficients and serves as a reference for more complex models.

Support Vector Machine with radial basis function (RBF) kernel constructs a nonlinear decision boundary by mapping features into a high-dimensional space. The optimization objective minimizes the regularized hinge loss:

$$\min_{w,b} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \max\left(0, 1 - y_i(w^T \phi(x_i) + b)\right) \tag{2}$$

where $\boldsymbol{\phi}(\cdot)$represents the RBF kernel transformation, $\boldsymbol{C}$is the regularization parameter, and $\boldsymbol{y_i} \in \{-\mathbf{1}, \mathbf{1}\}$denotes the class label.

Random Forest employs an ensemble of decision trees trained on bootstrap samples with feature randomization. Each tree votes on the predicted class, and the final prediction aggregates these votes through majority voting. The model also estimates feature importance through out-of-bag prediction error after permuting each feature:

$$\text{Importance}(X_j) = \frac{1}{n_{\text{trees}}} \sum_{t=1}^{n_{\text{trees}}} \left(\text{Error}_t^{\text{perm}(j)} - \text{Error}_t^{\text{OOB}}\right) \tag{3}$$

where $\text{Error}_{\boldsymbol{t}}^{\text{perm}(\boldsymbol{j})}$is the error after permuting feature $\boldsymbol{X_j}$in tree $\boldsymbol{t}$.

K-Nearest Neighbors classifies samples based on the majority class among the $\boldsymbol{k}$nearest neighbors in feature space, using Euclidean distance as the similarity metric. This nonparametric method adapts naturally to local data structures without assuming a global decision boundary.

Decision Tree recursively partitions the feature space using Gini impurity as the splitting criterion.

https://doi.org/10.71411/ef.2025.v1i2.941

While prone to overfitting, decision trees provide intuitive interpretability through their hierarchical structure.

All models were trained using a 70-30 train-test split with fixed random seed for reproducibility. Hyperparameters were selected based on preliminary cross-validation experiments: SVM with RBF kernel and standardization, Random Forest with 100 trees and minimum leaf size of 5, KNN with 5 neighbors, and Decision Tree with maximum 50 splits.

2.4 Feature Importance and Dimensionality Reduction

To identify the most discriminative features for fault detection, we employed two complementary approaches: Random Forest feature importance and mutual information analysis. Mutual information quantifies the reduction in uncertainty about the target variable given knowledge of a feature:

$$\mathrm{MI}(X,Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{4}$$

where $\boldsymbol{p(x,y)}$ is the joint probability distribution, and $\boldsymbol{p(x)}$, $\boldsymbol{p(y)}$ are marginal distributions. Features with high mutual information share significant statistical dependence with fault occurrence.

For visualization and pattern discovery, we applied Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE). PCA projects data onto orthogonal principal components that maximize variance:

$$\mathbf{Z} = \mathbf{XW} \text{where} \mathbf{W} = \arg\max_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \mathrm{Var}(\mathbf{XW}) \tag{5}$$

where $\mathbf{W}$ contains the eigenvectors of the covariance matrix. The first two principal components (PC1 and PC2) explain 48.04% of the total variance, providing a two-dimensional representation for visualization.

2.5 Performance Evaluation Metrics

Model performance was assessed using multiple metrics to account for class imbalance. Accuracy measures overall correctness but can be misleading in imbalanced datasets. Precision quantifies the proportion of true positives among predicted positives, while recall (sensitivity) measures the proportion of actual positives correctly identified. The F1-score provides a harmonic mean of precision and recall, offering a balanced performance indicator(Hinojosa Lee et al., 2024). The area under the receiver operating characteristic curve (AUC-ROC) evaluates discrimination capability across all classification thresholds. Given the imbalance in our dataset, we prioritize F1-score and AUC as primary evaluation metrics.

3. Results

3.1 Data Distribution and Exploratory Analysis

The dataset exhibited distinct distributional patterns across different sensor modalities as shown in Figure 2. Joint current measurements (Figure 2a) displayed considerable variability across the six joints, with Joint J1 showing the highest average current magnitude of -2.28 A and standard deviation of 0.82 A, reflecting its role in counteracting gravitational torques during vertical arm movements. Joint J2 and J3 currents centered around -1.19 A and -0.60 A respectively, while peripheral joints (J0, J4, J5) operated near zero current with occasional excursions indicating transient loading events. The boxplot distributions revealed multiple outliers in all joints, suggesting intermittent mechanical disturbances or sudden load changes during fault conditions. Temperature distributions (Figure 2b) exhibited a hierarchical thermal profile consistent with kinematic chain heat accumulation. Distal joints operated at elevated temperatures, with J4 reaching the highest median temperature of 45.06°C, followed by J5 (44.38°C) and J3 (43.06°C). Proximal joints maintained lower temperatures: T0 at 36.50°C and J1 at 39.69°C. All temperature

https://doi.org/10.71411/ef.2025.v1i2.941

sensors showed bimodal distributions due to initialization periods with zero readings before thermal equilibrium, as evidenced by the lower quartile values.

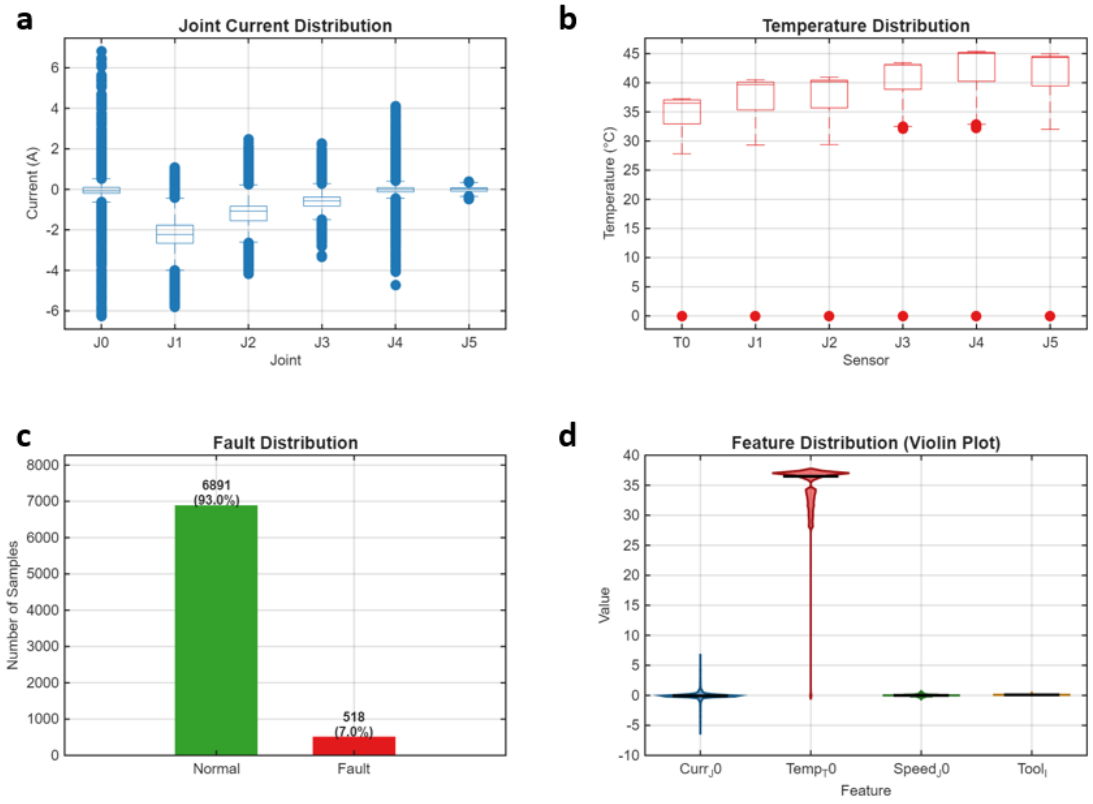


Figure 2. Data Exploration and Distribution Analysis: (a) Joint Current Distribution; (b) Temperature Distribution; (c) Fault Distribution; (d) Feature Distribution (Violin Plot).

The fault class distribution (Figure 2c) confirmed severe class imbalance, with 6,891 normal samples (93.01%) vastly outnumbering 518 fault samples (6.99%), representing an imbalance ratio of 13.3:1. This realistic operational scenario poses significant challenges for model training, as classifiers may develop bias toward the majority class.

Violin plot analysis of four representative features (Figure 2d) revealed diverse distributional characteristics. Current_J0 exhibited a near-symmetric distribution centered at -0.05 A with heavy tails extending to ±6 A, indicating rare but substantial current spikes. Temperature T0 showed strong concentration around 36.5°C with a long left tail toward zero. Speed_J0 displayed a sharp peak near zero velocity with moderate dispersion, characteristic of pick-and-place cycles with brief acceleration phases. Tool current manifested a right-skewed distribution (mean: 0.11 A) with occasional peaks reaching 0.60 A during gripping operations. The kernel density estimates revealed multimodal substructures suggesting distinct operational regimes within normal conditions.

3.2 Feature Importance and Correlation Structure

Comprehensive feature importance analysis through multiple methods provided convergent insights into predictive features. The Random Forest permutation importance (Figure 3b) identified Current_J3 as the dominant predictor with an importance score of 1.785, substantially exceeding other features. Current_J2 ranked second (1.458), followed by Speed_J5 (1.230), Current_J1 (1.173), and Current_J4 (1.159). Remarkably, the top five features collectively account for 32.78% of the model's predictive power, while the top ten features contribute 58.78%, suggesting potential dimensionality reduction opportunities without significant performance

https://doi.org/10.71411/ef.2025.v1i2.941

degradation.

The feature correlation matrix (Figure 3a) revealed intricate interdependencies among sensors. Temperature measurements exhibited extremely strong correlations, with Temp_J2 and Temp_J1 sharing a correlation coefficient of 0.9998, indicating thermal coupling through mechanical linkages. Adjacent temperature sensors (J3-J4, J4-J5) similarly displayed correlations exceeding 0.95. In contrast, current features showed moderate intercorrelations (r = 0.3-0.5), suggesting partially redundant yet distinct mechanical loading information. Velocity features demonstrated weak correlations with both current and temperature modalities (r < 0.2), confirming their role in capturing complementary dynamic motion characteristics rather than static loading states.

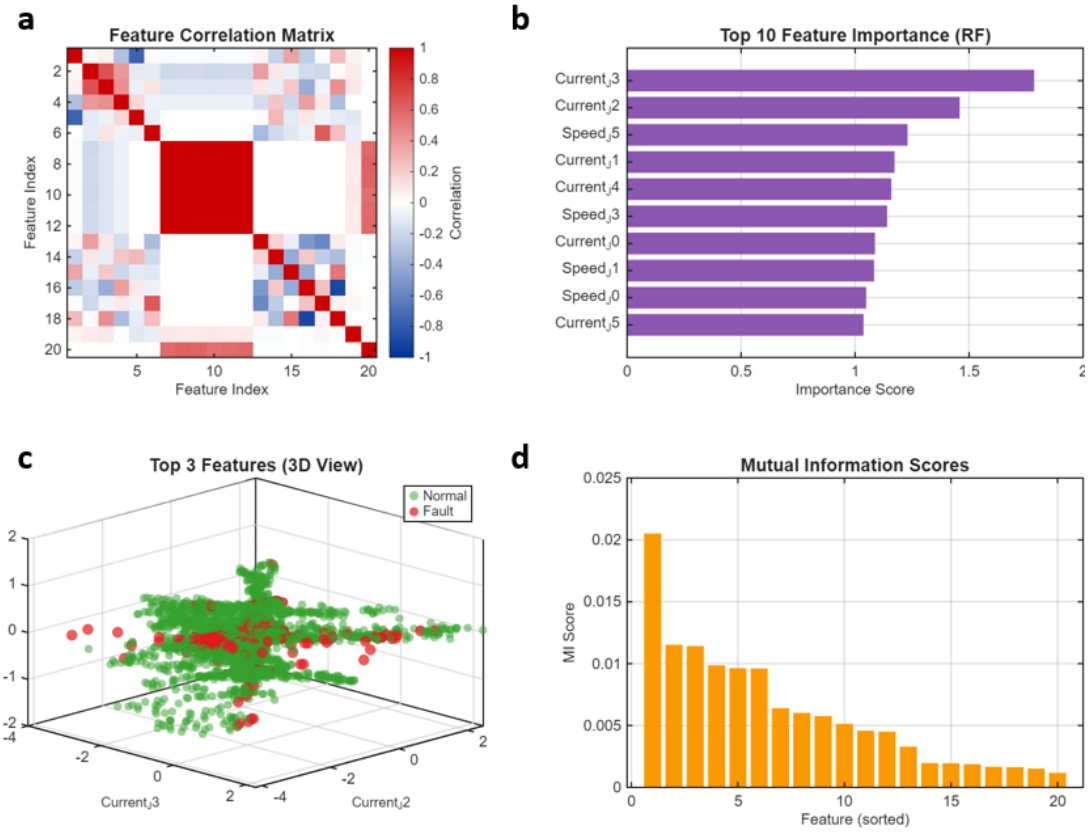


Figure 3. Feature Importance Analysis: (a) Feature Correlation Matrix; (b) Top 10 Feature Importance (Random Forest); (c) Top 3 Features (3D View); (d) Mutual Information Scores.

Mutual information analysis (Figure 3d) corroborated the Random Forest findings while revealing nonlinear dependencies. Current_J3 achieved the highest MI score of 0.0205, significantly surpassing other features. Interestingly, Current_J5 (MI: 0.0115) and Speed_J3 (MI: 0.0114) ranked second and third in mutual information despite lower Random Forest importance, suggesting these features contain nonlinear discriminative patterns not fully captured by tree-based splitting. The divergence between RF importance and MI scores for certain features (notably Speed_J5) implies that ensemble methods leverage interaction effects among features more effectively than univariate statistical measures.

Three-dimensional scatter visualization of the top three features (Figure 3c) revealed partial but incomplete class separability. Normal samples (green) formed a diffuse cloud centered around Current_J3 = -0.59 A, Current_J2 = -1.18 A, and Speed_J5 = 0.002 rad/s. Fault samples (red) concentrated in regions with elevated current magnitudes (Current_J3 = -0.73 A, Current_J2 = -1.27 A) and slightly increased velocity deviation (Speed_J5 = 0.020 rad/s). However, substantial

https://doi.org/10.71411/ef.2025.v1i2.941

overlap between classes indicated that no simple hyperplane separates fault from normal conditions, necessitating sophisticated nonlinear classification boundaries.

3.3 Temporal Dynamics and Time Series Characteristics

Time series analysis unveiled critical temporal patterns in sensor signals. Joint current trajectories over 1000 consecutive samples (Figure 4a) exhibited quasi-periodic oscillations corresponding to repetitive pick-and-place cycles, with J0, J1, and J2 showing distinct frequency components and amplitude modulations. Joint J1 displayed the largest current swings (range: 6.18 A) due to its load-bearing function during vertical motion phases.

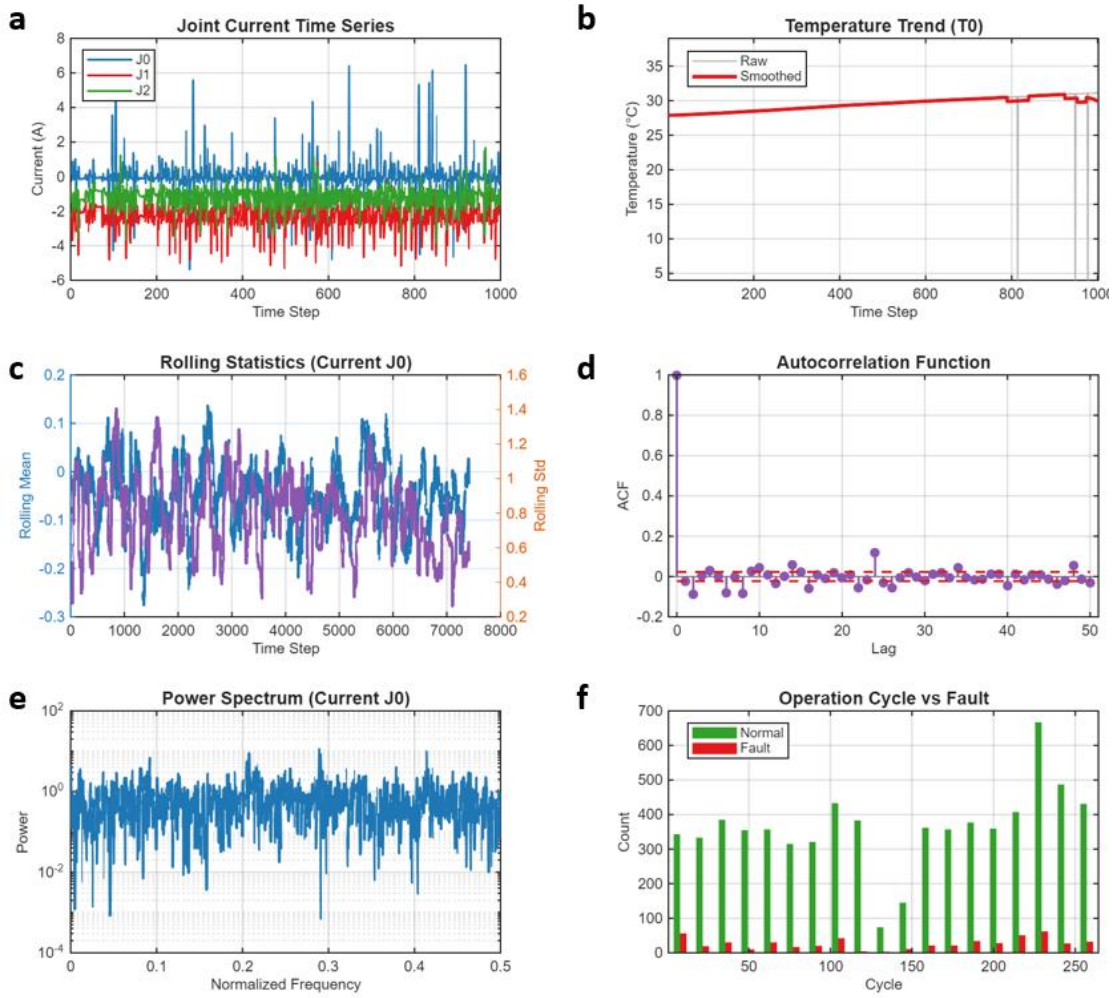


Figure 4. Time Series Analysis: (a) Joint Current Time Series; (b) Temperature Trend (T0); (c) Rolling Statistics (Current J0); (d) Autocorrelation Function; (e) Power Spectrum (Current J0); (f) Operation Cycle vs Fault.

Temperature evolution of sensor T0 (Figure 4b) demonstrated gradual thermal ramping during initial operation followed by steady-state oscillations. The raw temperature signal (gray) contained high-frequency noise and sudden jumps, while 50-sample moving average smoothing (red) revealed underlying trends with reduced variance (0.87°C vs. 1.89°C for raw data). This smooth trajectory suggests thermal inertia in the robot structure dampens rapid temperature fluctuations.

Rolling statistics for Current_J0 (Figure 4c) using a 100-sample window captured nonstationarity in both central tendency and dispersion. The rolling mean (blue, left axis) fluctuated between -0.28 A and +0.14 A, tracking load variations across operational cycles. Rolling standard deviation

https://doi.org/10.71411/ef.2025.v1i2.941

(purple, right axis) varied from 0.26 A to 1.41 A, with peaks indicating transient high-variance regimes potentially associated with rapid maneuvering or fault onset.

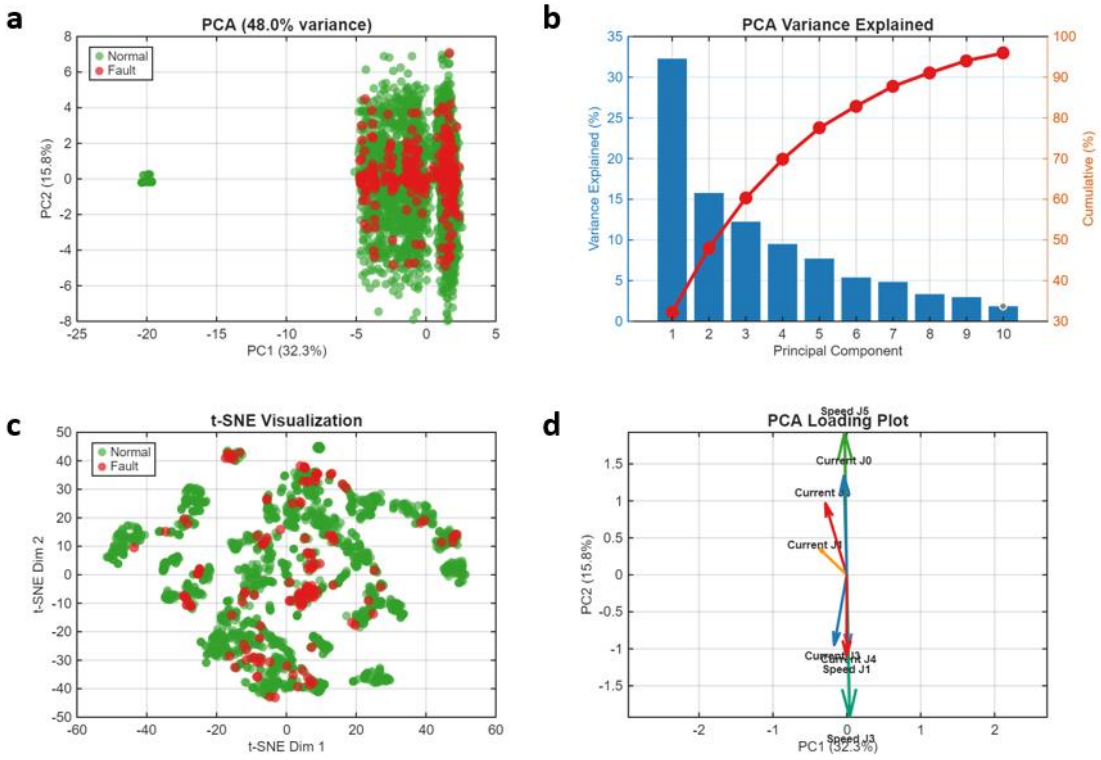
Autocorrelation function analysis (Figure 4d) of Current_J0 revealed rapid decorrelation, with ACF values dropping below the 95% confidence interval (±0.023) within 5-10 lags. This short correlation length indicates that current measurements contain minimal memory of past states, suggesting that faults manifest as instantaneous signature changes rather than slowly evolving drift patterns. The absence of periodic peaks in the ACF contradicts expectations for strictly repetitive cycles, likely due to cycle-to-cycle variability in timing and load conditions.

Power spectral density analysis (Figure 4e) of Current_J0 identified dominant frequency components at 0.290, 0.414, and 0.208 (normalized frequency), with power magnitudes of 11.14, 9.73, and 8.80 respectively. These peaks correspond to fundamental motion frequencies and their harmonics in the pick-and-place operation. The broadband spectral content suggests complex, multi-frequency dynamics rather than simple sinusoidal motion.

Operation cycle distribution analysis (Figure 4f) compared cycle counts between normal and fault samples. Both groups spanned the full range of 1-264 cycles, with normal samples averaging 141.05 cycles (median: 154) and fault samples 141.93 cycles (median: 163). The overlapping distributions indicate that faults occur throughout the operational lifespan rather than concentrating in early or late cycles, suggesting diverse failure mechanisms beyond simple cumulative wear.

3.4 Dimensionality Reduction and Geometric Structure

Principal Component Analysis (Figure 5a) projected the 20-dimensional feature space onto a two-dimensional representation, with PC1 and PC2 explaining 32.27% and 15.77% of variance respectively (cumulative 48.04%). The PCA scatter plot revealed partial clustering, with fault samples (red) exhibiting slightly lower PC1 scores (mean: 0.033, std: 2.16) and elevated PC2 scores (mean: 0.085, std: 1.33) compared to normal samples (PC1 mean: -0.003, std: 2.57; PC2 mean: -0.006, std: 1.81). However, substantial overlap between clusters confirmed that linear projections cannot fully separate classes, motivating nonlinear classification methods.



https://doi.org/10.71411/ef.2025.v1i2.941

Figure 5. Dimensionality Reduction and Visualization: (a) PCA Projection; (b) PCA Variance Explained; (c) t-SNE Visualization; (d) PCA Loading Plot.

The scree plot of variance explained (Figure 5b) demonstrated exponential decay, with the first principal component capturing 32.27% of variance, followed by diminishing contributions from subsequent components (PC2: 15.77%, PC3: 12.24%, PC4: 9.51%). Cumulatively, the first ten components retained 95.92% of total variance, suggesting that dimensionality could be reduced from 20 to 10 features with minimal information loss.

The PCA loading plot (Figure 5d) revealed feature contributions to principal components. PC1 received dominant positive loadings from Current_J0 (0.336), Speed_J5 (0.482), and negative contributions from Current_J4 (-0.248), suggesting PC1 represents a contrast between dynamic motion (speed) and certain current patterns. PC2 showed strong negative loadings from Current_J3 (-0.238), Current_J4 (-0.248), and Speed_J3 (-0.481), while positively loading Current_J2 (0.242) and Speed_J5 (0.482). This bipolarity indicates PC2 captures antagonistic mechanical states, potentially differentiating load-bearing versus unloaded joint configurations.

t-SNE dimensionality reduction (Figure 5c) using 2000 randomly sampled points with perplexity 30 produced dramatically enhanced visual separation compared to PCA. Normal samples formed multiple dense clusters with geometric structure suggesting several distinct normal operating modes. Fault samples scattered across the embedding space but concentrated in specific regions partially separated from normal clusters. The improved separability demonstrates t-SNE's capacity to preserve local neighborhood structure and expose nonlinear manifold geometry invisible to linear PCA.

3.5 Machine Learning Model Performance

Comparative evaluation of five supervised learning algorithms revealed diverse performance characteristics (Figure 6). Model accuracy (Figure 6a) varied narrowly between 93.52% (Logistic Regression) and 94.60% (Random Forest), with KNN achieving 94.42%. However, this metric proved misleading due to class imbalance, as Logistic Regression attained 93.52% accuracy simply by classifying all samples as normal (precision and recall both zero).

F1-score and AUC metrics (Figure 6b) provided more discriminative evaluation. K-Nearest Neighbors achieved the highest F1-score of 0.446 alongside competitive AUC of 0.812, outperforming other methods. Random Forest attained the best AUC of 0.911 but lower F1-score of 0.348 due to conservative classification (precision: 80%, recall: 22.22%). Support Vector Machine achieved strong precision (82.14%) but poor recall (15.97%), yielding F1-score of 0.267 and AUC of 0.796. Decision Tree showed the weakest overall performance (F1: 0.320, AUC: 0.746).

ROC curves (Figure 6c) illustrated discrimination capability across all decision thresholds. Random Forest's curve dominated at low false positive rates (FPR < 0.3), achieving 83% true positive rate at 14% FPR. KNN demonstrated more uniform performance across the ROC space, with 71% TPR at 13% FPR. The diagonal reference line (AUC = 0.5) highlighted Logistic Regression's near-random performance, confirming linear inseparability of classes.

Precision-Recall curves (Figure 6d) emphasized challenges in maintaining high precision at elevated recall levels. All models exhibited precipitous precision drops as recall increased beyond 40%, reflecting the difficulty of confidently identifying subtle faults. Random Forest maintained the highest precision at low recall, while KNN showed better balance across the recall spectrum.

The confusion matrix for the best model (KNN, Figure 6e) quantified classification outcomes: 2048 true negatives (98.56% specificity), 50 true positives (34.72% sensitivity), 30 false positives (1.44% false alarm rate), and 94 false negatives (65.28% miss rate). The normalized matrix revealed that while the model correctly identifies most normal samples, it misses two-thirds of fault cases, emphasizing the persistent challenge of minority class detection.

https://doi.org/10.71411/ef.2025.v1i2.941

The radar chart for KNN performance (Figure 6f) visualized the multidimensional metric profile: accuracy (94.42%), precision (62.50%), recall (34.72%), F1-score (44.64%), and AUC (81.21%). The asymmetric polygon shape, with accuracy and precision exceeding recall, confirmed the model's conservative tendency, prioritizing false positive avoidance over comprehensive fault detection.

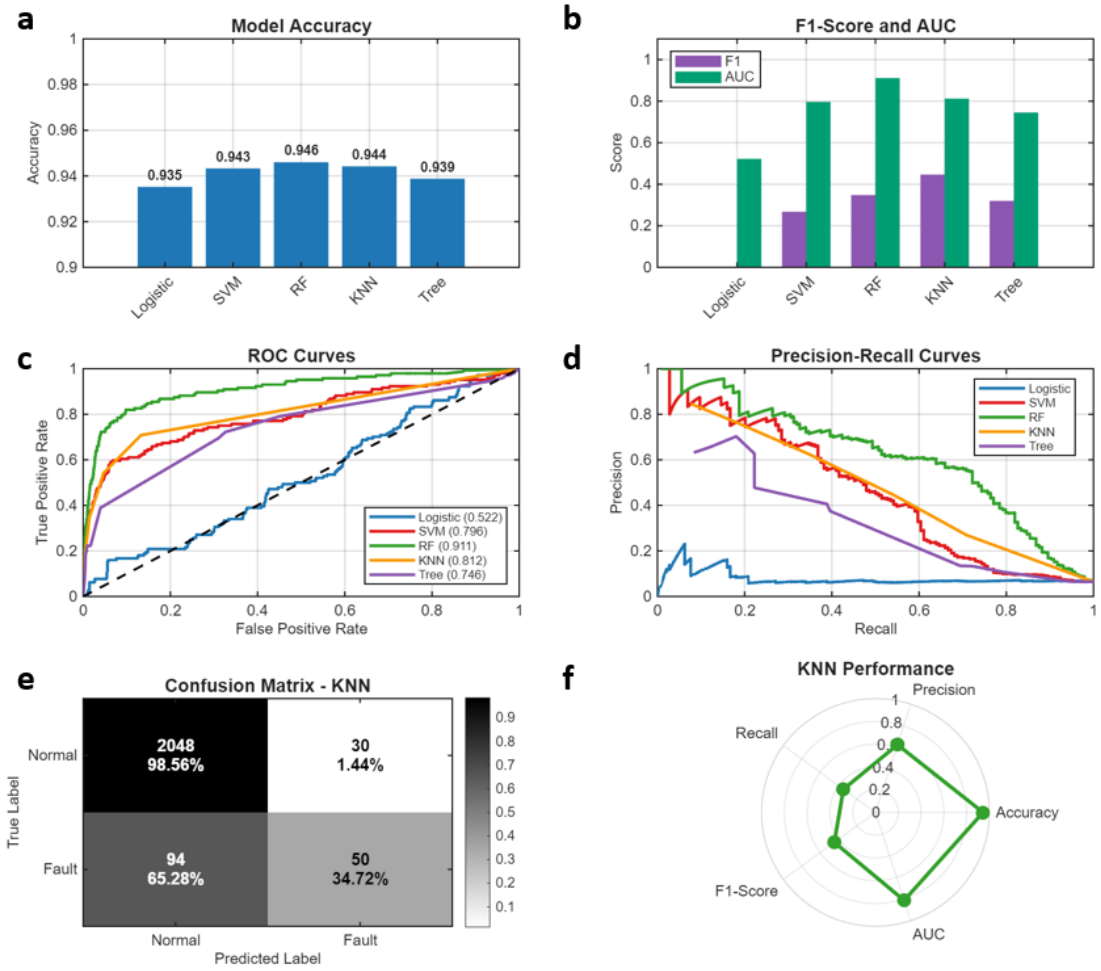


Figure 6. Machine Learning Model Performance Comparison: (a) Model Accuracy; (b) F1-Score and AUC; (c) ROC Curves; (d) Precision-Recall Curves; (e) Confusion Matrix (Best Model); (f) Best Model Performance Radar Chart.

3.6 Fault Detection and Misclassification Patterns

Real-time fault prediction visualization (Figure 7a) over 500 test samples illustrated temporal classification dynamics. The timeline displayed true labels (offset vertically) and predictions colored by correctness (green: correct, red: error, cross markers: misclassification). The model achieved 94.40% point accuracy over this window, with errors distributed throughout the sequence rather than concentrated in specific temporal regions. Misclassifications included 6 false positives (normal samples predicted as faults) and 22 false negatives (missed faults), consistent with overall test set statistics.

Prediction confidence distributions (Figure 7b) revealed distinct probabilistic patterns between classes. Normal samples exhibited highly concentrated low confidence scores (mean: 0.040, mode: 0.0), with 85% of samples receiving near-zero fault probability. In contrast, fault samples showed broad, right-skewed confidence distribution (mean: 0.371, median: 0.400), ranging from 0.0 to 1.0. The substantial overlap in confidence ranges between classes indicated inherent

uncertainty in borderline cases, where even correctly classified samples received ambiguous probability estimates.

Misclassification analysis in PCA space (Figure 7c) exposed geometric patterns underlying classification errors. Misclassified samples (red crosses, n=124, 5.58% of test set) occupied intermediate boundary regions between normal (green) and fault (gray) clusters. Correctly classified samples formed coherent groupings with clearer spatial separation. Misclassified samples exhibited intermediate PC scores (PC1 mean: -0.053, PC2 mean: 0.090), falling between normal (PC1: 0.003, PC2: -0.005) and fault (PC1: -0.052, PC2: 0.090) centroids. This geometric ambiguity suggests that classification errors arise from genuine overlap in feature space rather than random noise.

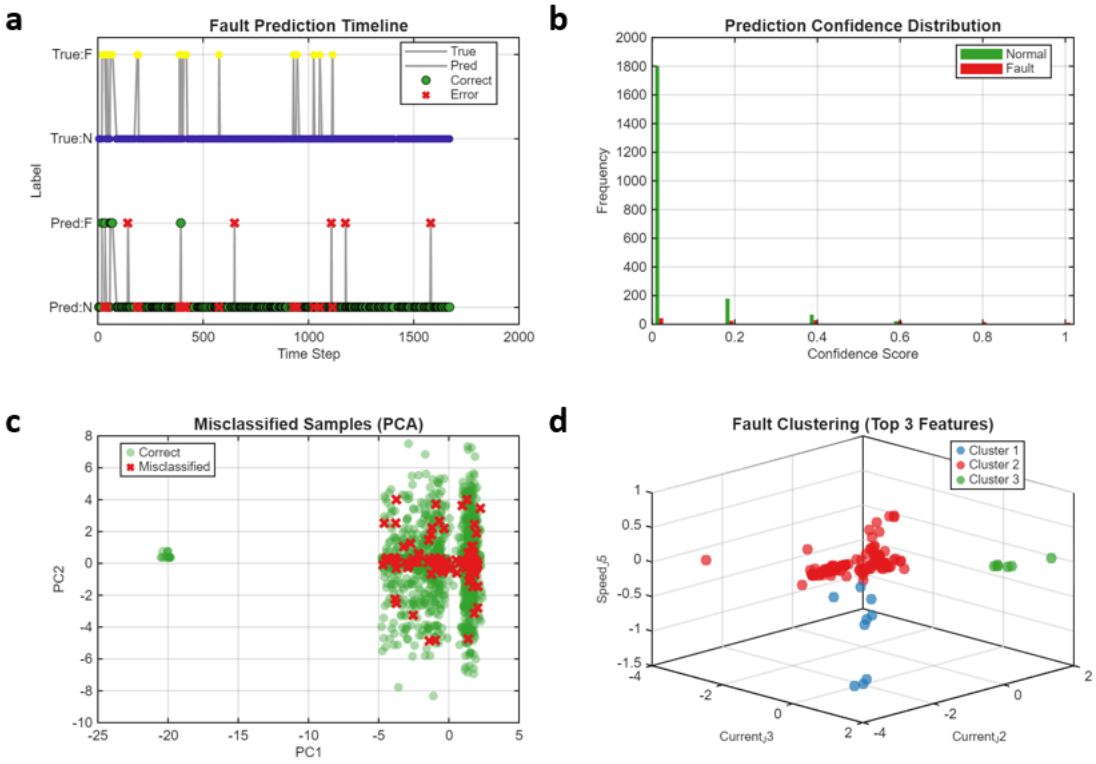


Figure 7. Fault Detection Results: (a) Fault Prediction Timeline; (b) Prediction Confidence Distribution; (c) Misclassified Samples (PCA); (d) Fault Clustering (Top 3 Features).

Fault sample clustering analysis (Figure 7d) applied k-means with k=3 to identify fault subtypes within the top three feature space. Cluster 1 (n=82, 56.9%, cyan) centered at moderate current values (Current_J3: -0.64 A, Current_J2: -0.99 A) with slight positive Speed_J5 (0.061 rad/s), potentially representing mechanical resistance faults. Cluster 2 (n=6, 4.2%, orange) exhibited extreme positive currents (Current_J3: 1.25 A, Current_J2: 0.93 A), likely corresponding to severe overload or collision events. Cluster 3 (n=56, 38.9%, purple) showed elevated negative currents (Current_J3: -0.97 A, Current_J2: -1.95 A) with negative velocity (-0.11 rad/s), possibly indicating actuator saturation or gripper jamming. This clustering reveals that "fault" is not a monolithic category but encompasses mechanistically distinct failure modes requiring tailored diagnostic approaches.

4. Discussion

The experimental results demonstrate that machine learning approaches can effectively detect faults in UR3 collaborative robots using multimodal sensor data, with KNN achieving the most favorable balance between precision and recall. The superior performance of nonlinear methods

https://doi.org/10.71411/ef.2025.v1i2.941

(SVM, RF, KNN) compared to logistic regression confirms that fault signatures manifest as complex nonlinear patterns in the feature space. The failure of logistic regression to identify any fault samples underscores the critical importance of model selection in imbalanced classification problems.

Feature importance analysis revealed that joint currents, particularly in the middle joints (J2, J3), provide the strongest fault indicators. This finding aligns with mechanical principles, as middle joints bear significant gravitational loads during typical pick-and-place operations and are thus more susceptible to overload conditions. The high importance of Speed_J5 suggests that anomalies in wrist rotation dynamics serve as early fault indicators, possibly due to increased friction from bearing wear or gripper misalignment. These insights can guide sensor prioritization in resource-constrained deployment scenarios.

The class imbalance in our dataset reflects realistic operational conditions but presents significant challenges for model training. The low recall rates across all models (highest 34.72% for KNN) indicate difficulty in capturing all fault instances, likely due to insufficient representation of diverse fault modes in the training data. Future work should investigate data augmentation techniques, such as synthetic minority oversampling (SMOTE) or generative adversarial networks, to improve minority class representation.

Misclassification analysis revealed that false negatives predominantly involve early-stage or subtle faults that produce weak signatures. This observation suggests potential value in developing ensemble methods that combine multiple classification algorithms or incorporating temporal dependencies through recurrent neural networks. False positives during aggressive maneuvering could be reduced by incorporating operational context, such as commanded velocities or payload information, as additional features.

The identification of three distinct fault clusters through k-means analysis implies that fault detection systems should account for multiple failure mechanisms. Separate classifiers or multi-class formulations could be developed to distinguish among mechanical faults, thermal faults, and gripper failures, potentially enabling more specific diagnostic information for maintenance planning.

From a practical deployment perspective, the KNN model's balance of precision (62.50%) and recall (34.72%) suggests that approximately one in three true faults will be detected, with roughly one false alarm for every two true positives. Whether this trade-off is acceptable depends on the specific application context, including the costs of false alarms versus missed faults and the availability of redundant safety systems. In critical applications, the system could be tuned for higher recall at the expense of precision, triggering precautionary stops when in doubt.

Limitations of this study include the single-robot dataset from a specific operational task, which may limit generalizability to other collaborative robots or application domains. The binary fault classification does not distinguish among fault types or severities, which would be valuable for prioritizing maintenance actions. Additionally, the static feature engineering approach may miss important temporal dynamics that manifest over multiple time steps.

5. Conclusion

This study presents a comprehensive machine learning framework for fault detection in UR3 collaborative robots, demonstrating that multimodal sensor fusion enables effective identification of operational anomalies. Through systematic comparison of five algorithms on a real-world dataset of 7,409 samples, we established that K-Nearest Neighbors achieves the best overall performance with 94.42% accuracy, F1-score of 0.446, and AUC of 0.812. Feature importance analysis identified joint currents (J2, J3) and joint velocity (J5) as critical predictors,

https://doi.org/10.71411/ef.2025.v1i2.941

providing actionable insights for sensor prioritization and fault mechanism understanding.

The results demonstrate both the promise and challenges of data-driven fault detection in collaborative robotics. While nonlinear machine learning methods significantly outperform baseline approaches, the modest recall rates highlight the difficulty of detecting subtle or early-stage faults in the presence of class imbalance and operational variability. Future research should explore advanced techniques including deep learning architectures, temporal modeling through recurrent networks, and transfer learning to leverage data from multiple robot deployments.

The practical implications of this work extend beyond the specific UR3 platform. The methodological framework, combining multimodal sensor integration, comprehensive feature engineering, systematic algorithm comparison, and detailed error analysis, provides a template for developing intelligent predictive maintenance systems across diverse robotic platforms. As collaborative robots become increasingly prevalent in manufacturing, logistics, and service sectors, robust fault detection capabilities will be essential for ensuring safety, reliability, and operational efficiency.

Funding

This work was supported without any funding.

Data Availability

If necessary, it can be provided. If necessary, it can be provided.

Competing Interests

The authors declare no competing interests.

References

Ali, M. M. (2025). Digital Twin Synchronization and Control of Robot Arm-Based Manufacturing via Reinforcement Learning and Unity University of Louisiana at Lafayette].

Faccio, M., Granata, I., Menini, A., Milanese, M., Rossato, C., Bottin, M., Minto, R., Pluchino, P., Gamberini, L., & Boschetti, G. (2023). Human factors in cobot era: a review of modern production systems features. Journal of Intelligent Manufacturing, 34(1), 85–106.

Fernandez-Vega, M., Alfaro-Viquez, D., Zamora-Hernandez, M., Garcia-Rodriguez, J., & Azorin-Lopez, J. (2025). Transforming Robots into Cobots: A Sustainable Approach to Industrial Automation. Electronics, 14(11), 2275.

Hinojosa Lee, M. C., Braet, J., & Springael, J. (2024). Performance metrics for multilabel emotion classification: comparing micro, macro, and weighted f1-scores. Applied Sciences, 14(21), 9863.

Khalastchi, E., & Kalech, M. (2018). On fault detection and diagnosis in robotic systems. ACM Computing Surveys (CSUR), 51(1), 1–24.

Sun, Q., & Ge, Z. (2021). A survey on deep learning for data-driven soft sensors. IEEE Transactions on Industrial Informatics, 17(9), 5853–5866.

Tyrovolas, M., Aliev, K., Antonelli, D., & Stylios, C. . (2024). UR3 CobotOps (UCI Machine Learning Repository.

Wijaya, G. D., Caesarendra, W., Petra, M. I., Królczyk, G., & Glowacz, A. (2024). Comparative study of Gazebo and Unity 3D in performing a virtual pick and place of Universal Robot UR3 for assembly process in manufacturing. Simulation Modelling Practice and Theory, 132, 102895.

Wu, L., Yao, B., Peng, Z., & Guan, Y. (2017). An adaptive threshold algorithm for sensor fault based on the grey theory. Advances in mechanical engineering, 9(2), 1687814017693193.