Data-Driven Regression Approaches for Early Anomaly Detection in Medium-Voltage Electric Motors

Nejc Friškovec
Elmont, d.o.o. Krško
Cesta krških žrtev 135E, Krško, Slovenia
Nejc.Friskovec@elmont-kk-si

Klemen Grozina
Nuklearna elektrarna Krško, d.o.o.
Vrbina 12, Krško, Slovenia
klemen.grozina@nek.si

Manja Obreza
Gen energija, d.o.o.
Vrbina 17, Krško, Slovenia
Manja.Obreza@gen-energija.si

Dalibor Igrec
University of Maribor, Faculty of Energy Technology
Hočevarjev trg 1, Krško, Slovenia
Dalibor.igrec@um.si

ABSTRACT

Reliable operation of medium-voltage motors is one of key prerequisites for availability of different equipment in nuclear power plants. Even with robust motor design margins, gradual degradation due to bearing wear, insulation ageing, reduced cooling performance and mechanical imbalance is still observed in these machines. Early symptoms often appear only as small variations in temperature, vibration or electrical load, which systems and procedures with fixed alarm thresholds typically do not detect. Possibility of more advanced diagnostic approach was studied. For several critical motors, ten years of measurements are available, including winding and bearing temperatures, vibrations, currents and relevant system parameters. The reconstruction of normal operating behaviour and the identification of deviations linked to developing faults are possible using this data. Before modelling can be performed, the dataset is pre-processed by removing corrupted records, interpolating missing values and normalizing variables. To enhance early fault detection, regression-based anomaly-detection models are applied. The models are trained exclusively on periods representing healthy operation, learning the relationships between environmental, system and motor-specific variables. Once trained, models generate expected values for key parameters such as bearing temperature. Deviations between predicted and measured values serve as indicators of abnormal behaviour, particularly when residuals show persistent shifts, exceed statistical limits or trigger cumulative deviation metrics. A portion of the historical dataset is withheld for validation to assess model accuracy on previously unseen data. Metrics such as the coefficient of determination and root mean square error are used to quantify predictive performance, while fault detection capability is evaluated on periods containing simulated fault conditions. The results show that regression models can detect subtle anomalies significantly earlier than traditional threshold-based systems, providing

maintenance teams with valuable time to plan inspections or adjust monitoring strategies. The integration of regression models for early warning detection is regarded as a promising enhancement to maintenance practices and is considered to support the long-term reliability of critical rotating components.

Keywords: *Medium-Voltage Induction Motor, Predictive Maintenance, Regression Model, Neural Network, PCA*

1 INTRODUCTION

Maintenance of equipment in nuclear power plants is considered essential for ensuring safe and reliable operations. Medium-voltage induction motors are considered vital components, and their failures should be minimized. Historically a time-based preventive maintenance strategy is employed on majority of components, whereby equipment is inspected and serviced at fixed time intervals, irrespective of its actual condition [1, 2]. Main disadvantage of time-based approach is that maintenance periods are usually very conservative and equipment downtime due to maintenance and cost of overhaul may be a challenge. In some cases, conditional based approach which is based on fixed thresholds (e.g., a bearing temperature limit, vibrations limit), is implemented. That approach typically extends periods between overhauls. Although such simple condition-based approach is effective in preventing obvious failures, it cannot detect and predict some slow drifts and precursors of failures. These challenges are shifting attention from simple condition-based maintenance toward predictive maintenance, where early fault detection and optimized maintenance actions are emphasized.

Artificial intelligence (AI) and machine learning (ML) tools are being used in nuclear and other industries to analyse large operational datasets, through which hidden patterns of wear or abnormal behaviour can be identified [3, 4]. Their use has been widely demonstrated across various fields. In industrial applications, AI has been attributed broad potential, ranging from operational improvements and system design to predictive diagnostics. Evidence has shown that data-driven diagnostics can substantially improve the early detection of degradation, thereby reduce unnecessary work and costs, and enhance safe operation. Supervised learning (e.g., regression and classification) is applied when labelled fault histories are available. In contrast, unsupervised learning is employed to reveal patterns in unlabelled data and is particularly valuable when large fleets of sensors are involved. The increasing deployment of ML techniques has been reported across numerous fields for automatically detecting anomalies [5, 6]. This enables engineers to focus on the most significant deviations, thereby accelerating response and enhancing diagnostic precision in electric-motor applications.

The faster adoption of these solutions is being increasingly observed in the industry, accompanied by growing regulatory recognition that appropriate assessment and certification will be required [7-9]. The U.S. Nuclear Regulatory Commission [7] has issued its Artificial Intelligence Strategic Plan for 2023–2027, in which the first licensing applications incorporating AI are anticipated within a few years, and preparations of staff and processes for such reviews are outlined. Support for AI research has also been expressed by the International Atomic Energy Agency, which has established working groups on the technical and regulatory aspects of plant deployment. The need for transparent, explainable algorithms and robust validation frameworks before widespread use is consistently emphasized by experts, as trust in model outputs is regarded as paramount in safety-critical settings. Despite these challenges, existing studies indicate that deep learning, ensemble models, and hybrid methods can be successfully applied for fault detection in power plants, with associated enhancements in operational safety and efficiency being reported.

Within this context, a hybrid data-driven regression framework combining principal component analysis (PCA) and a neural network is applied to detect deviations in medium-voltage induction motors operating in a nuclear power plant. The use of multiple algorithms enables a comparative assessment of each approach. The objective of this introductory section is to demonstrate how AI

tools, when combined with extensive operational data, can be utilized to enhance the condition monitoring of medium-voltage motors and to identify deviations before they escalate into serious failures or outages. In this way, support can be provided for compliance with the Maintenance Rule (10 CFR 50.65) on monitoring maintenance effectiveness, while optimizing the reliability and availability of safety-significant plant components.

1.1 Main contributions

This paper makes several important contributions:

[1] A practical and reproducible framework for residual-based fault detection in medium-voltage induction motors used in nuclear power plants is presented. The framework integrates industrial data preprocessing, dimensionality reduction, nonlinear regression modelling, and statistical anomaly detection into a consistent workflow validated using long-term operational data.

[2] A robust data preparation methodology tailored to industrial time-series measurements is developed. The workflow includes automated quality control, redundancy handling, regular time-grid reconstruction, gap interpolation, and causal feature engineering, ensuring traceability and suitability for machine learning applications.

[3] A hybrid data-driven modelling approach based on principal component analysis and an artificial neural network is implemented for temperature prediction. The approach effectively captures nonlinear dependencies among environmental, operational, and machine-specific variables while reducing redundancy and improving numerical stability.

[4] A residual-based fault detection strategy is established using statistically defined thresholds derived from receiver operating characteristic (ROC) analysis. The proposed method demonstrates reliable detection of abnormal behaviour through interpretable deviation metrics.

[5] The study demonstrates the applicability of data-driven diagnostics in safety-critical environments by emphasizing transparency, validation discipline, and compatibility with condition-based maintenance practices in nuclear power plants.

2 TECHNICAL BACKGROUND

2.1 Faults in medium-voltage motors

Medium-voltage induction motors in nuclear power plants are widely used as prime movers for safety and non-safety-significant auxiliary systems. Although these machines are designed with substantial safety margins, gradual degradation of mechanical and electrical components remains unavoidable during long-term operation [10]. Among common failure modes, bearing-related faults represent the most frequent cause of degradation in medium-voltage motors [11, 12]. Bearings are exposed to combined mechanical, thermal, and lubrication stresses, particularly during transient operating conditions such as starts and load variations. Typical failure mechanisms include wear, surface damage, lubrication loss, misalignment effects, and thermal distress. Because many of these processes evolve gradually, early symptoms often appear as small but persistent changes in temperature behaviour rather than as abrupt failures [13].

Temperature monitoring is therefore widely recognized as a key diagnostic indicator for rotating machinery [14]. Unlike vibration-based indicators, which often respond strongly in advanced stages of degradation, temperature signals can reflect earlier stages of fault development. However, these changes are typically subtle and can be masked by environmental influences, operational variability and measurement noise. As a result, simple condition-based monitoring approaches relying on fixed alarm thresholds, such as temperature or vibration limits may fail to detect early-stage degradation and slow drifts. To overcome these limitations, data-driven diagnostics approaches have increasingly been applied in rotating equipment monitoring [15, 16]. In such approaches, models

are trained to represent normal operating behaviour using historical operational data, while deviations between measured and predicted values are interpreted as indicators of abnormal conditions. This concept forms the basis for predictive maintenance strategies, where early fault detection and optimized maintenance actions are emphasized. Such approaches are particularly suitable for complex industrial systems, where multiple correlated variables influence thermal behaviour and where explicit physical modelling is often impractical.

2.2 Data-driven diagnostics – workflow process

The predictive framework developed in this study combined statistical preprocessing, dimensionality reduction, and nonlinear regression using an artificial neural network. The objective was to construct a stable data-driven representation of normal operating behaviour suitable for residual-based fault detection. Because the available dataset contained numerous correlated variables from multiple sensor channels, preprocessing steps were applied to ensure robustness and numerical stability. Input variables were standardized using statistics computed exclusively from the training subset to prevent information leakage. To improve robustness against extreme values, standardized inputs were clipped using percentile-based limits (0.5th and 99.5th percentiles) derived from the training data. To address redundancy and multicollinearity, PCA was applied to the standardized inputs. The first 30 principal components were retained and used as model inputs, preserving dominant system dynamics while reducing dimensionality and noise sensitivity. The resulting PCA scores were additionally standardized using training-set statistics to ensure stable neural-network training.

On the reduced feature space, a feed-forward artificial neural network was trained to model the nonlinear relationship between operational inputs and the target temperature. The multilayer perceptron architecture consisted of two hidden layers with 64 and 32 neurons, respectively, using rectified linear unit activation functions. Dropout layers with a rate of 0.1 were applied after each hidden layer to improve generalization. The output layer consisted of a single neuron followed by a regression layer. Training was performed using the Adaptive moment estimation optimizer with an initial learning rate of 1×10^{-3}, a mini-batch size of 512, and a maximum of 300 batches. L2 regularization of 1×10^{-5} and early stopping with a validation patience of 10 were used to reduce overfitting. The dataset was split chronologically, with the first 80% used for training and the remaining 20% for independent validation.

Residual analysis formed an integral part of the methodology. The residual, defined as the difference between measured and predicted temperature, provided a physically interpretable indicator of abnormal behaviour. Persistent increases in residual magnitude were interpreted as deviations from learned normal operation. The resulting framework provides a computationally efficient and interpretable approach suitable for industrial monitoring applications. The whole workflow construction is presented in Fig. 1

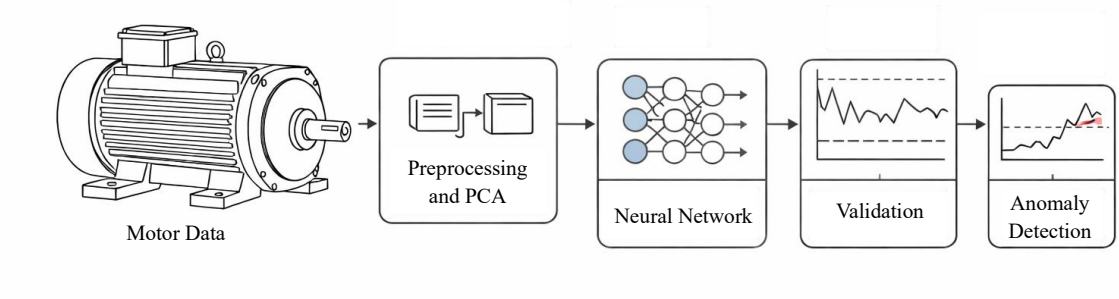

Figure 1: Workflow construction.

3 DATA PREPARATION AND ANALYSIS

3.1 Correlation and dependency analysis

Correlation and dependency analysis was performed to obtain an initial understanding of relationships among measured variables and to identify dominant drivers of bearing temperature. Because industrial time-series data often contain shared operating influences, correlation was interpreted as an indicator of association rather than causation. Both Pearson and Spearman coefficients were used to capture linear and monotonic relationships. Fig. 2 shows the relationship between ambient temperature and upper guide bearing temperature. The narrow upward-sloping scatter indicates a strong monotonic dependency, confirming ambient temperature as a key driver. At the same time, the spread around the trend line suggests that additional factors, such as loading conditions and cooling performance, contribute to variability.

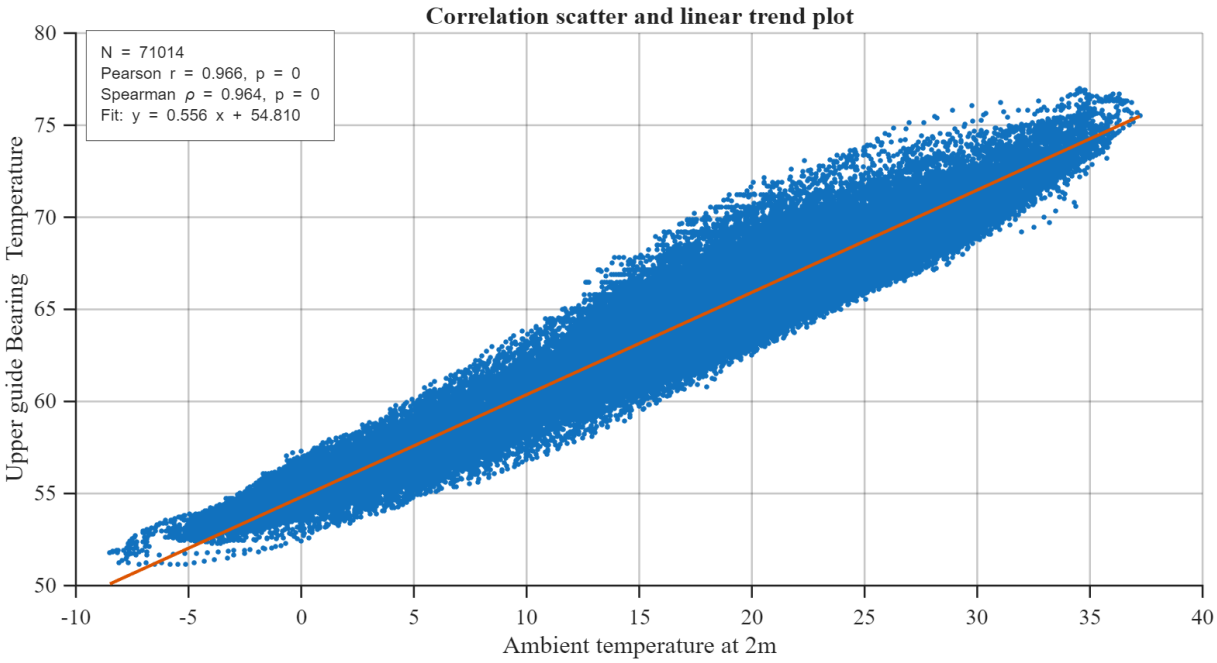

Figure 2: Ambient temperature vs upper guide bearing temperature.

To illustrate joint dependencies, a three-dimensional visualization was used in Fig. 3, where river temperature is plotted against ambient and bearing temperatures. The fitted plane indicates that a large portion of variability can be explained by shared seasonal effects, while deviations highlight additional influences not captured by pairwise analysis.

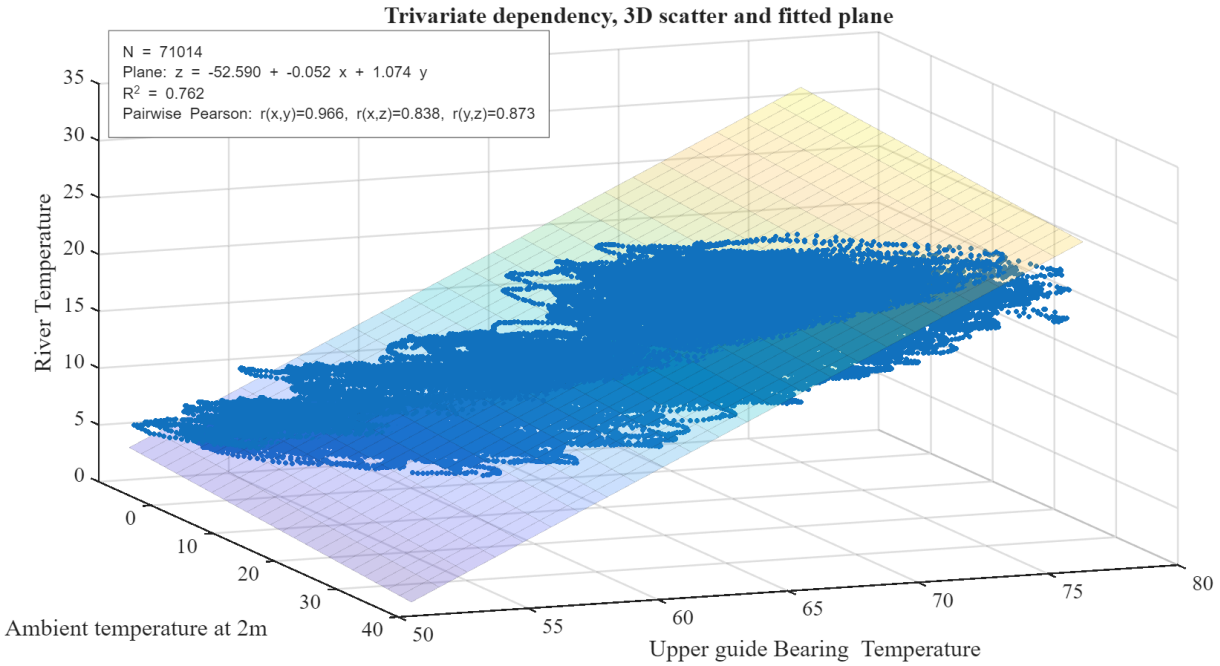

Figure 3: Bearing temperature dependency on ambient and river temperature.

Because the final predictive models operate in a higher-dimensional feature space, correlation analysis was used primarily as an exploratory step to identify dominant variables and motivate later dimensionality reduction rather than as a standalone modelling tool.

3.2 Data pre-processing

A structured preprocessing workflow was implemented to convert raw operational measurements into a consistent modelling dataset. Data was imported from multi-sensor time-series records sampled at approximately ten-minute intervals. After parsing timestamps, invalid and duplicate records were removed, and the data were mapped onto a regular time grid to ensure consistent alignment across variables. An operational filter was applied to retain periods corresponding to stable system operation, reducing confounding effects from mixed regimes. Redundant sensor channels were automatically merged, and inconsistent readings were excluded through conflict detection rules. Quality control focused primarily on temperature signals and included plausibility limits, step-change constraints, moving-median outlier detection, and stuck-sensor identification. Short missing segments were interpolated using shape-preserving methods, while longer gaps were retained as missing. The impact of preprocessing on each temperature channel is summarized in Fig. 4, confirming that most data remained measured rather than reconstructed.

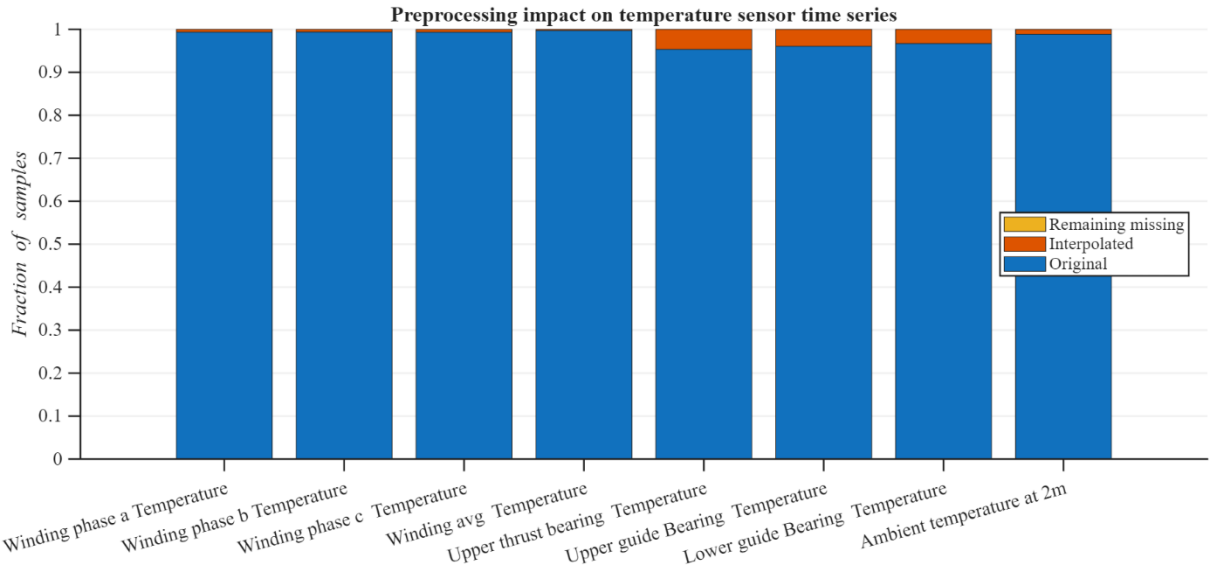

Figure 4: Preprocessing impact on temperature sensor time series.

The final dataset was constructed by removing remaining missing values and generating causal time-domain features, including lags, differences, and trailing statistics. These steps provided a robust and traceable input representation for subsequent machine learning modelling.

4 MODEL TRAINING AND EVALVATION

The predictive model for the upper guide bearing temperature was developed using a data-driven regression approach based on operational time-series measurements. The modelling objective was to reconstruct normal thermal behaviour and enable fault detection through residual analysis. The input feature space consisted of selected temperature signals, environmental variables, process measurements, and engineered time-domain features. Feature engineering included lagged variables, first-order differences, and trailing moving statistics designed to capture thermal inertia and transient operating effects. Because strong correlations were present among inputs, PCA was applied after standardization to reduce dimensionality and improve numerical stability. Only leading components explaining most of the variance were retained for modelling.

On the reduced feature space, a feed-forward artificial neural network was trained to approximate the nonlinear relationship between operational variables and the target temperature. The model was trained using standardized inputs, mini-batch gradient-based optimization, and regularization to ensure stable convergence and prevent overfitting. Model performance was evaluated using chronological data split, where the first 80% of observations were used for training and the remaining 20% for independent validation. The trained model demonstrated strong agreement with measured data, as shown in Fig. 5, confirming its ability to reproduce dominant thermal dynamics and provide a reliable baseline for residual-based fault detection.

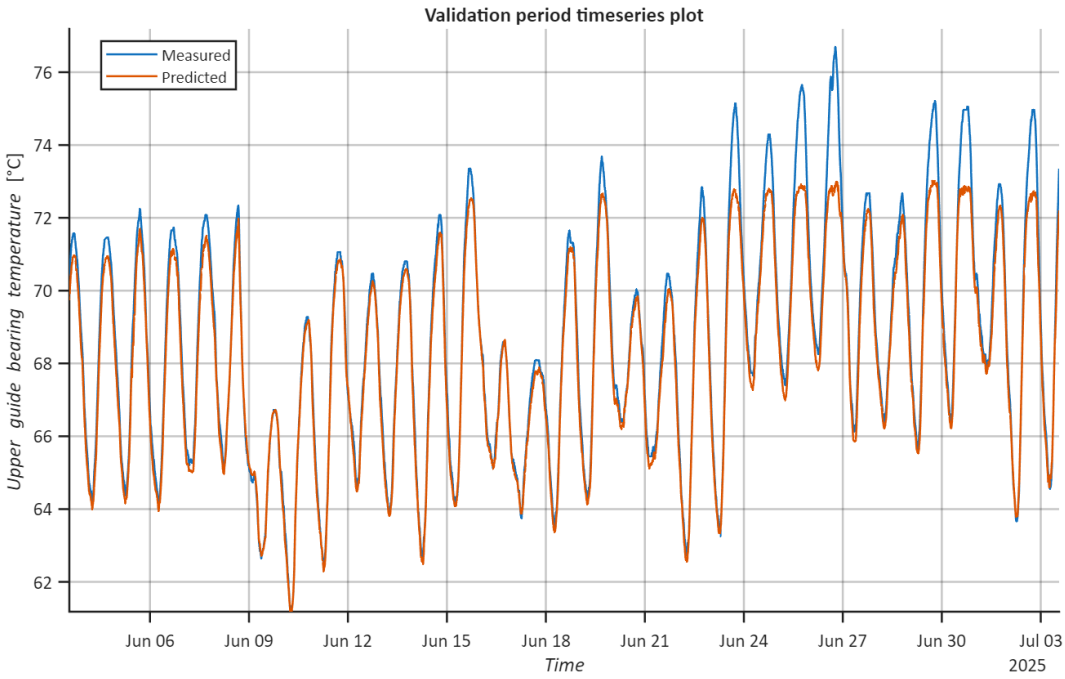

Figure 5: Model prediction vs. measured data.

4.1 Validation

Model validation represents a critical step in assessing its suitability for practical fault detection applications. Because operational datasets are inherently time-dependent, a chronological split was used to evaluate generalization performance. The first 80% of the dataset was used for training, while the remaining 20% was reserved for independent validation. This time-based partition prevents information leakage from future samples and provides a realistic assessment of performance under real operating conditions.

Several standard regression metrics were used to quantify model performance. The coefficient of determination was used to measure the proportion of output variance explained by the model. Values of 0.9945 for the training set and 0.9464 for the validation set indicate excellent explanatory capability and stable generalization. The root mean square error was 0.368 °C for training and 0.692 °C for validation, demonstrating low prediction error relative to the operating temperature range.

Additional performance indicators further confirmed model reliability. The mean absolute error and median absolute error remained low in both datasets, indicating consistent performance without large deviations. The relative metrics mean absolute percentage error and symmetric mean absolute percentage error also showed small values, confirming that prediction errors remained small relative to absolute temperature levels. The mean bias error was close to zero, indicating minimal systematic deviation.

4.2 Fault detection

Fault detection was performed using residual analysis, where the residual was defined as the difference between measured and predicted temperatures. Because deviations are relevant regardless of sign, the absolute residual was used as the anomaly score. This approach relies on the assumption that under normal operating conditions the residual remains small and approximately stationary, while faults produce systematic deviations that the model cannot reproduce.

The fault detection threshold was determined using ROC analysis. The absolute residual was treated as a classification score and compared with known fault labels obtained from simulated fault

conditions. For each possible threshold, the true positive rate (TPR) and false positive rate (FPR) were calculated, and the ROC curve was constructed. The optimal threshold was selected using the Youden index, which maximizes the difference between TPR and FPR and provides a balanced trade-off between detection sensitivity and false alarms.

Fig. 6 shows the time evolution of the residual signal, where a clear increase is observed at the onset of the fault. The residual remains consistently above the selected threshold during faulty operation, enabling reliable detection.

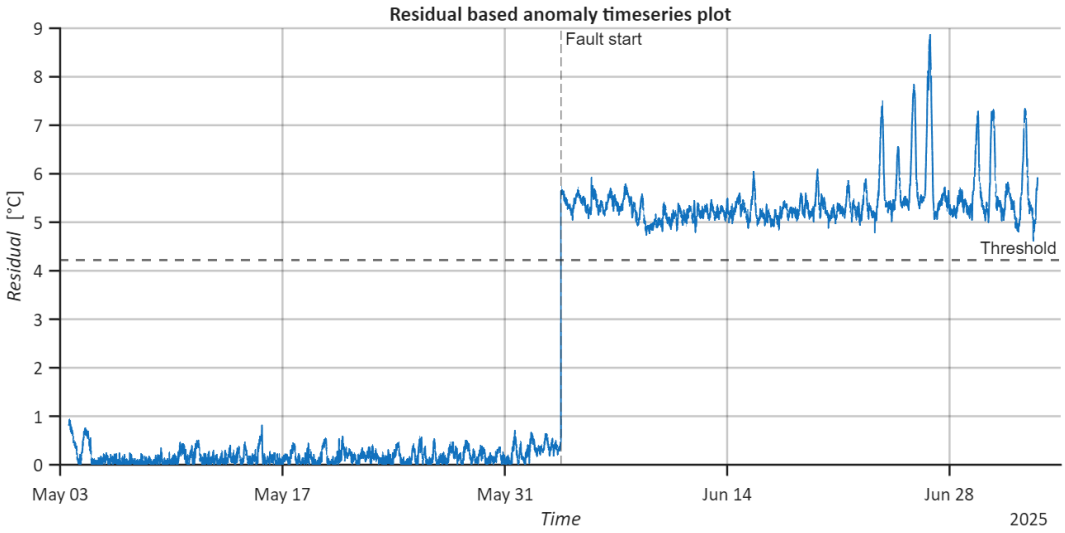

Figure 6: Residual based anomaly detection at fault occurrence.

Fig. 7 presents the corresponding confusion matrix, demonstrating clear separation between normal and faulty states in the analysed scenario.

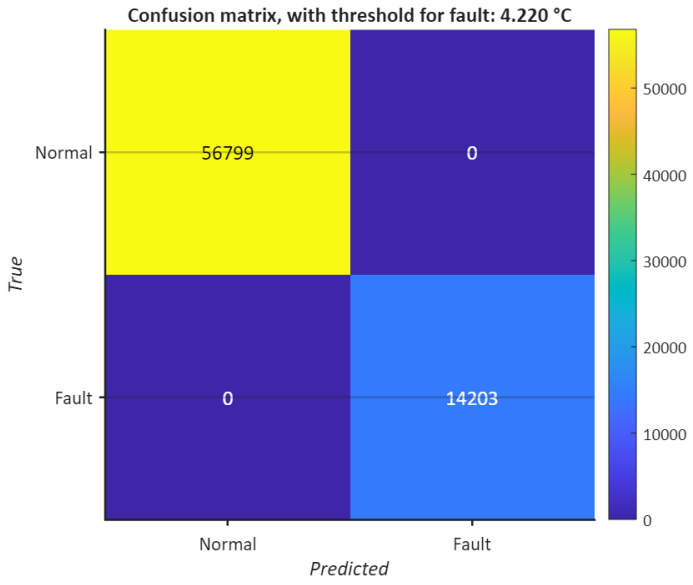

Figure 7: Confusion matrix.

This residual-based approach provides a physically meaningful detection mechanism because faults alter the thermal equilibrium of the system in ways that cannot be captured by a model trained in normal operation [6]. Consequently, the residual increases and serves as a direct indicator of abnormal behaviour. The main advantage of this method lies in its robustness and interpretability, as

the detection result directly reflects deviations from expected physical behavior rather than relying on opaque classification features.

5 FUTURE WORK

In future work, the proposed methodology is expected to be extended toward broader generalization and practical deployment across multiple identical machines operating under varying conditions. Although the current study demonstrates reliable fault detection using a model trained on a single unit, further improvements can be achieved by incorporating operational data from several equivalent motors during the training phase. Such an approach would allow unit-to-unit variability to be captured explicitly, thereby improving robustness and enabling the model to distinguish between normal structural differences and genuine fault-related deviations. In addition, limited calibration procedures based on short healthy operating intervals are planned, through which systematic offsets between units can be compensated without retraining the full model.

Another important direction concerns the extension of the modelling framework to additional diagnostic variables. While this study focused on the upper guide bearing temperature, the same residual-based methodology can be applied to other critical parameters, including winding temperatures, lower bearing temperatures, and selected process variables. By developing a coordinated set of parameter-specific models, it is anticipated that earlier and more reliable detection of abnormal operating conditions can be achieved, while also supporting improved fault isolation through comparative residual analysis.

Further research will also address real-time implementation and adaptive monitoring strategies. The existing workflow is already compatible with online deployment, as it requires only fixed preprocessing transformations and a trained predictive model. However, future work will investigate adaptive thresholding approaches that account for seasonal variations, long-term system drift, and changing load conditions. Such extensions are expected to reduce false alarm rates while preserving sensitivity to early-stage degradation.

6 CONCLUSION

This study demonstrated the applicability of data-driven regression methods for early anomaly detection in medium-voltage induction motors operating in a nuclear power plant environment. A structured workflow integrating industrial time-series preprocessing, dimensionality reduction using principal component analysis, nonlinear regression based on an artificial neural network, and residual-based anomaly detection was developed and evaluated using long-term operational data.

The results showed that the proposed modelling approach can accurately reproduce dominant thermal dynamics across varying operating conditions. Validation performed using a chronological data split confirmed strong predictive performance and stable generalization, with low prediction errors relative to the operating temperature range. These findings indicate that the model provides a reliable representation of normal operation suitable for practical industrial monitoring applications. Residual-based evaluation further demonstrated that abnormal behaviour can be detected in a robust and interpretable manner. The use of statistically derived thresholds enabled consistent detection performance while maintaining low false alarm rates in the evaluated scenario. Because the detection mechanism is directly linked to deviations from expected behaviour, the approach remains transparent and appropriate for safety-critical industrial environments.

Overall, the results suggest that data-driven regression models represent a practical and scalable solution for improving diagnostic capabilities in medium-voltage motors and other critical rotating equipment, with clear potential to support predictive-based maintenance and enhance long-term operational reliability.

REFERENCES

[1] R. Han, P. Li and Z. Shi, “Implementation strategy of predictive maintenance in nuclear power plant,” in Proc. PHM-2022 London, London, UK, 2022, pp. 143-146.

[2] T. Herzog and K. Bartecki, “Predictive maintenance for electrical motors: Current approach and usage of artificial intelligence algorithms,” in Proc. MMAR, 2024, pp. 494-498.

[3] C. Lu et al., “Nuclear power plants with artificial intelligence in Industry 4.0 era,” IEEE Access, vol. 8, pp. 194315-194332, 2020.

[4] C. Tang et al., “Deep learning in nuclear industry: A survey,” Big Data Mining and Analytics, vol. 5, no. 2, pp. 140-160, 2022.

[5] S. Mari et al., “Machine learning for anomaly detection in induction motors,” MetroXRAINE, 2023.

[6] H. El Hadraoui et al., “Data-driven diagnostics for electric traction systems,” EUROCON, 2023.

[7] U.S. NRC, Artificial Intelligence Strategic Plan FY2023-2027, NUREG-2261, 2023.

[8] U.S. NRC, Exploring Advanced Computational Tools with AI/ML in Operating Nuclear Plants, NUREG/CR-7291, 2022.

[9] IAEA, Considerations for Deploying Artificial Intelligence Applications in the Nuclear Power Industry, Vienna, 2025.

[10] A. J. Bazurto et al., “Causes and failures classification of industrial electric motor,” IEEE ANDESCON, 2016.

[11] K. Kudelina et al., “Mechanical bearing faults for predictive maintenance,” IEEE SDEMPED, 2023.

[12] A. H. Bonnett, “Root cause failure analysis for AC induction motors,” IEEE PCIC, 2010.

[13] A. Siddique et al., “Applications of AI techniques for induction machine stator fault diagnostics,” IEEE SDEMPED, 2003.

[14] X. Liang, “Temperature estimation and vibration monitoring for induction motors,” IEEE EPEC, 2017.

[15] F. M. C. Santos et al., “Neural network classifier for faults detection in induction motors,” ICCAT, 2013.

[16] R. Yaqub et al., “Electrical motor fault detection using random forest classifier,” IEEE ASET, 2023.

[17] V.-N. Pham et al., “AI monitoring and diagnosing electric motor faults,” IEEE ICOIN, 2024.