Energy and AI 22 (2025) 100631

Contents lists available at ScienceDirect

Energy and AI

journal homepage: www.sciencedirect.com/journal/energy-and-ai

Data-driven machine learning model estimates efficiency gains from passive filters under variable loads

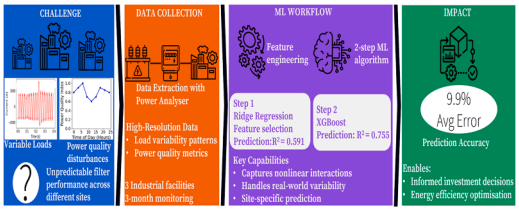
Uchenna Johnpaul Aniekwensi [a,*], Dipyaman Basu [b], Jörg Bausch [a]

[a] Institute of Sustainable Energy Systems (INES), Hochschule Offenburg, Badstraße 24, Offenburg 77652 Baden-Württemberg, Germany
[b] Livarsa GmbH, Im Fruchtfeld 17, Berghaupten 77791 Baden-Württemberg, Germany

HIGHLIGTS

- A two-step ML model predicts passive filter efficiency using field measurements.
- Leverages real-world data, overcoming simulation limits for site-specific estimates.
- Selected top features by Ridge Regression used by XGBoost to improve prediction.
- Neutral current, unbalance, harmonics, power factor: key drivers of efficiency gains.
- Enables data-driven investment for confident passive filter deployment.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Passive filters
Energy efficiency
Machine learning
Power quality
Predictive analytics
Data-driven model

ABSTRACT

Accurately estimating power loss reduction from passive filters before installation is challenging due to variable loads and power quality conditions across grid points. Existing studies rely on simulation or analytical models. These approaches often fail to capture real-world variability through data-driven methods. This gap limits effective, site-specific filter deployment decisions. We present a two-step machine learning approach to estimate energy efficiency gains from passive filters under variable conditions using high-resolution power analyzer data. Ridge Regression identifies key predictive variables, achieving baseline $R^2 = 0.591$. XGBoost then captures nonlinear interactions between load variability, power quality disturbances, and filter performance, improving accuracy to $R^2 = 0.755$. The methodology was validated through deployment at three industrial facilities in collaboration with Livarsa GmbH. Results demonstrate 9.9% average relative error across measured efficiency gains, confirming reliability under real-world conditions. Comprehensive validation through k-fold cross-validation, ensemble methods, and external testing quantified prediction uncertainty inherent in small industrial datasets (25 training samples). The approach offers a scalable, data-driven decision-support tool overcoming simulation-based limitations. Computational efficiency enables real-time assessment during client consultations without specialized software. Economic value derives from reduced performance guarantee margins, accelerated assessment timelines, and minimized warranty exposure. Limitations include statistical constraints from limited training data, reflected in cross-validation overfitting and wide confidence intervals. External validity requires site-specific validation for facilities with substantially different electrical characteristics. Despite these constraints, the findings provide practical value for energy professionals seeking efficient power quality solutions, enabling confident passive filter deployment decisions based on quantified performance predictions.

* Corresponding author.
E-mail address: uchenna.aniekwensi@hs-offenburg.de (U.J. Aniekwensi).

https://doi.org/10.1016/j.egyai.2025.100631

Available online 17 October 2025

1. Introduction

Accurately estimating energy efficiency benefits from passive filters before installation remains challenging. This difficulty stems from highly variable loads and power quality conditions across different grid points. Power quality issues, harmonic distortion, voltage unbalance, sag and swell, and high dynamic reactive power demand significantly influence energy efficiency and operational reliability [1]. The transition to non-linear, semiconductor-based loads has intensified these challenges, leading to increased energy losses and compromised equipment reliability [2].

Passive filters offer efficient, cost-effective solutions for improving electrical power quality in industrial installations. They minimize harmonics and distortions, increase power factor, and maximize economic benefits through direct energy savings while preventing equipment damage [3]. Under high-current loads, they are often the only viable option for power quality improvement. However, efficiency improvement potential varies greatly between installations due to unique load characteristics, baseline power quality, and operational patterns at each facility.

Existing approaches fail to provide reliable site-specific predictions. Most studies depend on simulation models to estimate efficiency gains via THD (Total Harmonic Distortion) reduction [4] or quantify efficiency potential through minimizing harmonic loss factor and maximizing loading capability [5]. Analytical models have been integrated into optimization algorithms to estimate potential through minimized active power losses [6]. These simulation-based and analytical methods share critical limitations: they fail to accurately reflect real-world variability and do not measure site-specific energy efficiency gains under actual operational conditions.

This gap creates significant commercial and technical challenges. Passive filter providers face performance guarantee risks and potential reputational damage from inaccurate predictions. System designers, operators, and end users require reliable filtration solutions optimized for their unique environments, yet traditional approaches cannot adequately address these needs. The lack of predictive tools tailored to operational dynamics limits effective filter deployment decisions and confidence in expected returns on investment.

The critical research gap lies in the absence of data-driven predictive tools that capture real-world operational variability for site-specific passive filter performance estimation. While simulation models provide theoretical insights, they cannot account for the complex interactions between variable loads, power quality disturbances, and filter performance under actual operating conditions. Data-driven approaches are essential for managing the dynamic complexities of modern electrical systems. Studies on smart grid optimization confirm that predictive modeling achieves quantifiable improvements in efficiency and peak demand reduction [7,8]. Improving energy efficiency through accurate prediction is critical for modern power systems, ensuring sustainable development while maintaining grid resilience and reliability.

We introduce a data-driven machine learning approach that addresses this gap by accurately estimating efficiency gains from passive filters using high-resolution power analyzer data collected from operational installations. Our two-step methodology employs Ridge Regression for feature importance ranking and dimensionality reduction, followed by XGBoost for nonlinear prediction. This approach precisely estimates efficiency gains under variable load and power quality conditions that cannot be captured by simulation.

The methodology involves thorough feature engineering on current waveforms, effective values, and aggregated power quality indices from real operational data. Ridge Regression handles multicollinearity among power quality variables and ranks feature importance through regularized coefficients, providing interpretable baseline predictions. XGBoost then captures complex nonlinear relationships between load variability, power quality disturbances, and filter performance that linear models cannot represent.

Our work builds on expanding research applying machine learning to energy savings prediction and power quality analysis [9,10]. While earlier studies examined Ridge Regression for energy applications [11, 12] and XGBoost for energy forecasting [13,14], our specific contribution targets efficiency improvements from passive filters in variable power quality environments using field measurement data. This addresses a critical gap between theoretical modeling and practical power systems engineering where operational variability dominates system behavior.

The practical implications are substantial. The approach provides a scalable tool supporting data-driven investment decisions in power quality improvement. By enabling accurate, site-specific predictions under real-world variability, it enhances passive filter planning and deployment confidence. Computational efficiency allows real-time assessment during client consultations without specialized simulation software. Economic value is derived from reduced performance guarantee margins, accelerated assessment timelines from weeks to minutes, and minimized warranty exposure.

Transparent acknowledgement of scope limitations strengthens confidence in reported capabilities. The 25-sample training dataset imposes statistical constraints reflected in cross-validation overfitting and wide confidence intervals. External validity requires site-specific validation for facilities with substantially different electrical characteristics. These limitations reflect fundamental challenges in industrial data collection where practical barriers like economic burden, limited partner availability, and safety constraints preclude extensive dataset expansion. Data availability constraints [15] and model sensitivity [16] restrict generalization capabilities and indicate potential result variability.

Despite these constraints, this work represents a significant advancement in applying machine learning for pragmatic power quality solutions. Research was conducted with industrial partner Livarsa GmbH, providing access to operational data from 28 facilities and enabling field validation through deployments at three test sites. This paper details methodology (data acquisition, feature engineering, model development, and validation) and validation results, demonstrating the efficacy of Ridge Regression and XGBoost in predicting energy savings while transparently addressing limitations and scope boundaries.

2. Related work

The increasing demand for energy efficiency and reliable power quality has spurred significant research into passive filter applications and the development of predictive models [17]. This section reviews relevant literature on power quality improvement, passive filter applications, and machine learning techniques for energy prediction, specifically focusing on how they address the challenges of variability in client-specific power quality scenarios.

2.1. Power quality and passive filter applications

The significance of maintaining power quality within acceptable limits, as defined by standards like IEEE 519, IEC 61000–3–6, and IEC 50160, has been widely studied. Harmonic distortion, voltage sags/swells, and power factor issues can greatly affect industrial equipment and energy efficiency [18]. Passive filters offer proven solutions for reducing harmonic distortion and enhancing power factor, including applications in three-phase grid-connected renewable energy systems [19]. Various design and optimization strategies for filters have been suggested, focusing on tuning methods and resonance suppression to achieve maximum harmonic attenuation [20,21]. Research into the theory of instantaneous power also influences the design and control of power filters [22]. While traditional studies often emphasize the theoretical advantages of passive filters, practical performance guarantees based on empirical, site-specific data are seldom addressed in the literature, highlighting the necessity for the predictive tools presented in this

work.

2.2. Machine learning for energy prediction and regression analysis

Machine learning techniques, particularly regression models, have increasingly been employed for energy prediction. Reviews such as [23] highlight the potential of regression models in capturing complex relationships for building energy consumption prediction. Ridge Regression, in particular, is valuable for handling multicollinearity and preventing overfitting, which are common issues in power quality data analysis, and its theoretical foundations and practical applications are discussed in [24]. However, the performance of these models often depends on the availability of sufficient data and the stability of training parameters.

Beyond handling multicollinearity and preventing overfitting using methods like Ridge Regression, machine learning algorithms offer powerful and promising solutions for energy efficiency in complex smart grids by predicting usage patterns, identifying inefficiencies, and supporting predictive maintenance. The demonstrated success of predictive modeling and optimization using algorithms including regression, clustering, and neural networks in achieving substantial benefits like 15 % energy efficiency improvement and 25 % peak demand reduction underscores ML's proven utility in industrial energy systems [7].

While established regression models like Ridge Regression are valuable for handling multicollinearity and preventing overfitting in power quality data, applying machine learning in decentralized or resource-constrained environments increasingly necessitates lightweight, scalable, and real-time solutions. For instance, novel frameworks utilizing Property Testing combined with Edge AI and Federated Learning are emerging to perform anomaly detection and classification efficiently in complex distributed networks (e.g., IoT), demonstrating significant reductions in processing time and energy consumption while maintaining high accuracy [25].

2.3. Tree-based ensemble methods: XGBoost and beyond

Beyond traditional regression models like Ridge Regression, tree-based ensemble methods such as Gradient Boosting and its optimized implementation, XGBoost [26], have demonstrated notable success in various prediction tasks, including energy forecasting [13,14]. These methods build a strong predictive model by combining the outputs of multiple decision trees, with each tree sequentially learning from the errors of its predecessors. XGBoost, in particular, integrates regularization techniques and efficient tree-building algorithms, often leading to superior performance compared to traditional linear models, especially when dealing with complex, non-linear relationships in the data. Studies have shown XGBoost' s effectiveness in capturing intricate patterns in energy consumption data and attaining high prediction accuracy [27]. Its capacity to handle non-linearities and feature interactions makes it a promising candidate for modelling the complex relationship between power quality parameters and efficiency gains. Furthermore, XGBoost provides mechanisms for assessing feature importance, offering insights into the drivers of the predicted outcomes [26]. While the initial phase of our work focused on the linear Ridge Regression model to establish a baseline and address multicollinearity, the potential to capture more complex relationships and achieve higher predictive accuracy motivated us to explore XGBoost as an alternative modelling approach. This shift aligns with the broader trend in machine learning towards utilizing powerful ensemble methods for challenging prediction problems.

The pursuit of optimal feature sets for complex prediction tasks has led to the adoption of advanced meta-heuristic optimization algorithms. These optimization methods are frequently employed for binary feature selection, aiming to improve predictive accuracy while minimizing feature complexity. For instance, specialized algorithms such as the Adaptive Dynamic Dipper Throated Optimization (bGW-DTO) have demonstrated superior ability in identifying minimal, high-impact feature subsets for classification problems, often achieving very high prediction accuracy and processing speed. Such sophisticated optimization approaches contrast with regularization techniques like Ridge Regression, demonstrating the expanding utility of feature selection across diverse domains like energy prediction and medical diagnosis [28].

The effectiveness of machine learning in complex, non-linear prediction problems often rely on robust model structure and hyperparameter tuning. Moving beyond generalized ensemble approaches, specialized methods utilize advanced meta-heuristic optimization algorithms such as the hybrid PSOBER (Particle Swarm Optimization and Al-Biruni Earth Radius Optimization) algorithm—to optimize complex network parameters (like those in CNNs and DBNs) for tasks requiring high predictive precision. These optimization techniques are crucial for maximizing accuracy and minimizing classification error in scenarios characterized by high data variability [29].

The complexity and dimensionality of data in industrial and network monitoring domains necessitate rigorous feature selection, particularly when traditional linear methods prove insufficient. Consequently, hybrid metaheuristic algorithms such as the GWDTO (Grey Wolf and Dipper Throated Optimization) method are increasingly utilized to solve the Feature Selection (FS) problem in high-dimensional domains like IoT network intrusion detection systems (IDS). These methods are designed to efficiently identify minimal, optimal subsets of features that maximize classification accuracy by enhancing the balance between the exploration and exploitation stages of the optimization process. This approach validates the general necessity of robust feature selection preceding high-accuracy prediction across diverse technical fields [30].

Furthermore, the optimization of complex ensemble models often relies on advanced techniques beyond traditional hyperparameter grids. Hybrid meta-heuristic optimization algorithms, such as the Adaptive Dynamic Sine Cosine Fitness Grey Wolf Optimization (ADSCFGWO) algorithm, have been successfully employed to optimize the weights and parameters of multi-model voting classifiers (NN, SVM, KNN) to boost accuracy significantly in complex non-linear prediction tasks like weed detection in drone imagery. The use of such optimization methods confirms the broader utility of meta-heuristics for enhancing ensemble learning performance in critical data-driven applications [31].

2.4. Addressing data limitations and model stability

A critical limitation in our study is the limited number of data samples, which can restrict the generalization capabilities of the model. Research on small sample size regression [15] emphasizes the challenges and potential biases associated with limited data. Beyond sample size constraints, ensuring external validity, confirming that the proposed methodology remains effective across diverse industrial sites outside the original validation set is a recurring limitation frequently raised in systematic literature reviews of machine learning frameworks. Techniques for mitigating these issues, such as regularization and cross-validation, are crucial for ensuring model robustness [32].

Furthermore, the observed sensitivity of our Ridge Regression-trained model to the random seed highlights the issue of model stability. Studies on random seed sensitivity in machine learning models [15, 16] explore the impact of initialization parameters on model performance. Techniques like ensemble learning and robust optimization can potentially address this limitation.

2.5. Novelty and contribution

Our work contributes significantly to the field by introducing a data-driven, two-step machine learning approach specifically tailored for accurately estimating power loss reduction from passive filters in variable power quality environments. This directly addresses the critical commercial and technical challenge of guaranteeing filter performance under real-world, dynamic conditions, which traditional simulation or

analytical models have largely failed to account for. While previous studies have explored the use of regression models for energy prediction, our approach uniquely focuses on quantifying passive filter efficiency improvements using high-resolution power analyzer data, encompassing rigorous feature engineering, and provides empirical validation with a focus on both linear (Ridge Regression for baseline and uncertainty) and non-linear (XGBoost for superior accuracy) modeling. The application of XGBoost, in particular, demonstrates a notable advance in capturing the complex, nonlinear interactions inherent in power quality data, offering a robust and scalable tool for data-driven investment decisions.

This novelty, which successfully bridges the gap between simulation-based studies and real-world predictive modeling for energy systems, demonstrates the strong practical impact achievable when applying advanced ML (Machine Learning) to specific domain challenges, akin to its effectiveness in complex diagnostic problems like Hepatitis C prediction in healthcare [8]. These predictive achievements in energy systems underscore ML's potential for increasing industrial decision-making and economic viability.

While our methodology utilizes Ridge Regression for feature selection and linear baseline establishment, current trends in data science increasingly incorporate meta-heuristic binary optimization methods (e.g., bADSCFGWO) to identify optimal feature subsets, thereby coupling advanced feature selection with optimized ensemble prediction models for superior overall performance and statistical significance. This shift towards dual-optimization validates the need for rigorous feature selection preceding non-linear prediction in diverse scientific domains [31].

Our two-step methodology, which uses Ridge Regression for reliable feature selection followed by XGBoost for capturing complex nonlinear interactions, directly addresses the challenge of guaranteeing passive filter performance under dynamic, real-world conditions. This emphasis on efficient feature extraction and robust non-linear prediction is paramount in overcoming resource and data limitations. This need for resource efficiency is further highlighted by cutting-edge solutions in related domains, such as the Property Testing and Federated Edge AI (Artificial Intelligence) framework, which achieves sublinear anomaly detection in constrained networks, showing the broad trend towards computationally frugal yet accurate data-driven decision-support tools [25].

3. Methodology

This section outlines the methodology used to develop a predictive machine learning model for estimating power loss reductions resulting from the installation of a passive filter. The process comprises data collection from various facilities, data processing, rigorous feature engineering, and the selection and training of the Ridge Regression and XGBoost algorithms. The overarching goal of the model is to predict the potential efficiency loss reductions for facilities before the filter installation. This prediction serves as a crucial input for assessing the economic viability of filter procurement based on the facility's existing power quality conditions. Fig. 1 provides a visual representation of the data flow during data collection for model training and its subsequent application for a new facility. To accurately estimate the filter's loss-reduction potential for a new facility, the model was trained using data captured when the passive filter was bypassed at existing facilities. The specific method for quantifying the efficiency gains (ECV®) is detailed in Section 3.2.1.

3.1. Data collection

The data acquisition process involved a measurement setup deployed at various facility locations, comprising a power analyzer [34] and a passive filter [35]. Critical electrical data, indicative of the power quality situation, were recorded at different time resolutions. Notably,

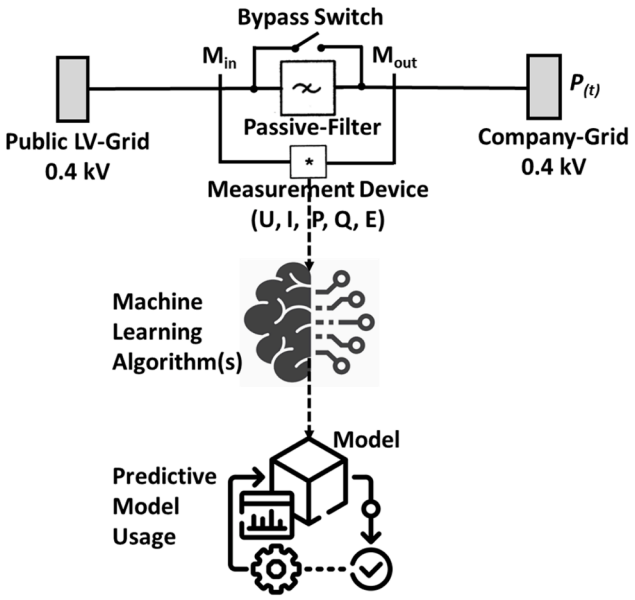


Fig. 1. Data capture and model usage workflow. A passive filter with a bypass switch is installed between the public LV (Low Voltage) transformer (0.4 kV) and the company grid (0.4 kV). Data captured and collected with a power analyzer are preprocessed and feature-engineered before training a machine learning model, which is then applied for decision-making [33].

the shutdown of the passive filter during the measurement campaign triggered a transient event, which was automatically recorded and stored within the power analyzer's web interface (accessible via the device's IP address). This event metadata was saved in PQDIF (Power Quality Data Interchange Format) and Comtrade (configuration and data) file formats.

The PQDIF files (IEEE Std 1159.3 compliant) served as the primary source for extracting two distinct data with different time resolutions using the PQDiffractor software [36]:

- 10-millisecond Effective Values: Representing half-cycle RMS values of current, voltage, etc., providing insights into the immediate impact of the passive filter bypass event.
- 40-microsecond Instantaneous Values: Capturing the detailed waveform characteristics, crucial for analyzing transient behavior and harmonic content.

Figs. 2 and 3 visually illustrate these two different time resolutions. Additionally, 2-second aggregated data [37] captured and stored centrally during the ECV® measurement period were utilized to provide a broader overview of the power quality conditions over time.

3.1.1. Dataset composition and industrial diversity

The dataset comprises 28 industrial facilities from Livarsa GmbH's client portfolio across Germany and Switzerland, representing seven distinct manufacturing sectors: plastic packaging, aluminum sand casting, closure technology, corrugated cardboard packaging, floor tile manufacturing, technical plastics systems (parts, assemblies, and technical systems), and building materials production (interior finishing and insulation). This sectoral diversity ensures representation of varied harmonic profiles, load characteristics, and power quality conditions typical across industrial applications.

Filter installation capacities range from 800A to 3200A, representing a fourfold variation in electrical infrastructure scale. This range captures both medium-scale manufacturing operations and large industrial facilities, providing applicability across typical commercial passive filter deployment scenarios. Each facility's measurements include comprehensive power quality data collected during normal operational

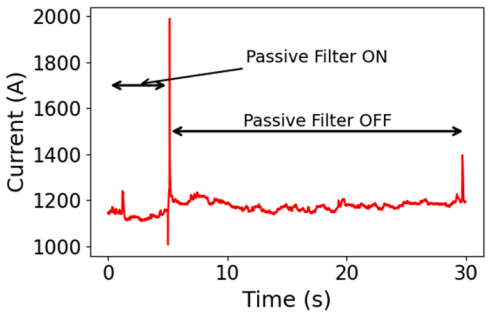


Fig. 2. Long-Term RMS Current Response to Filter Switching. The RMS current is shown versus time over 30 s. A large current transient occurs at $t \approx 5$s when the Passive Filter is switched from ON to OFF.

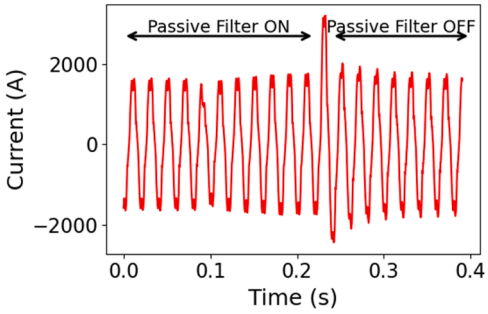


Fig. 3. AC (Alternating Current) Current Waveform During Filter Switching. The current is shown versus time over 400 ms. The Passive Filter is initially ON and is switched OFF at $t \approx 0.23$ s, causing a large transient spike followed by the continued, unfiltered AC current.

conditions, capturing realistic load variability and harmonic content.

While the absolute sample size of 28 reflects practical constraints in industrial data collection, the dataset's sectoral and scale diversity provides broader representativeness than laboratory or simulation-based studies with homogeneous conditions. The 25:3 training-testing split maintains this diversity across validation partitions.

3.2. Feature engineering

The feature engineering process involved extracting and transforming relevant information from the above-discussed raw data sources to create a comprehensive dataset for model training. This resulted in a total of 47 feature variables and the target variable (percentage efficiency gains) was calculated with the method described in the next section.

3.2.1. Target variable: efficiency gains data

The target variable, representing the efficiency gains achieved by the passive filter, was quantified using the ECV® (Energy Comparison Value) measurement method [38]. This method Fig. 4 is designed to provide a verifiable measure of energy efficiency across complex electrical networks, accounting for the non-uniform energy consumption patterns caused by fluctuating loads and operational conditions. The ECV® method facilitates a meaningful comparison of energy consumption with and without the filter in operation.

3.2.2. Feature variables

Feature variables were derived from the three primary data sources (10-millisecond effective values, 40-microsecond instantaneous values, and 2-second aggregated data).

3.2.2.1. Effective current data (10-millisecond resolution). During the 24-hour ECV® measurement, some PQDIF files (typically a file per hour during the main operating hours of the facility) are recorded. A sampled mean of the derived variables (e.g., S_p, Fig. 5) was calculated by averaging the values across the recorded observations within each of the PQDIF files collected during the ECV® measurement period using (1). Following a similar methodology to the calculation of S_p, other fundamental variables were engineered from the effective current data using Eqs. (2)-(6).

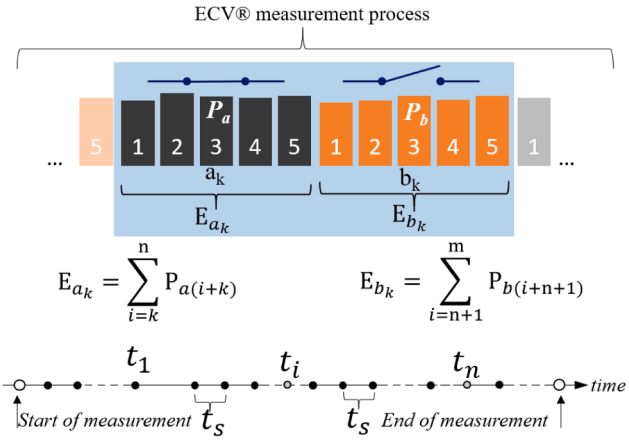


Fig. 4. Illustration of Energy Comparison Value Assessment. Intervals of energy measurements when the passive filter is on saving/bypass modes. The accompanying formulas define how the energy consumption for each interval is calculated by summing the corresponding recorded power measurement data over the respective measurement steps [38].

$$S_p = \mathfrak{S}_{i,t}^{N_{files},N_{samples}}\left[A_p(i,t)\right] \tag{1}$$

$$D_{p-q} = \mathfrak{S}_{i,t}^{N_{files},N_{samples}}\left[A_p,\ A_q(i,t)\right] \tag{2}$$

$$\mathfrak{a}(A)(i,t) = \ 1 \Big/ 3\sum_{p=1}^{3} A_p(i,t) \tag{3}$$

$$\mu = \mathfrak{S}_{i,t}^{N_{files},N_{samples}}\left[\mathfrak{a}(A)(i,t)\right] \tag{4}$$

$$\mathrm{DN}_p = S_p/\mu * 1000 \tag{5}$$

$$\mathrm{DN}_{p-q} = S_{p-q}/\mu * 1000 \tag{6}$$

$$\mathrm{DN\_N}_p = \mathfrak{a}\left(\mathrm{DN}_p\right)/\mu * 1000 \tag{7}$$

$$\mathrm{DN\_N}_{p-q} = \mathfrak{a}\left(\mathrm{DN}_{p-q}\right)/\mu * 1000 \tag{8}$$

where,

Indices:

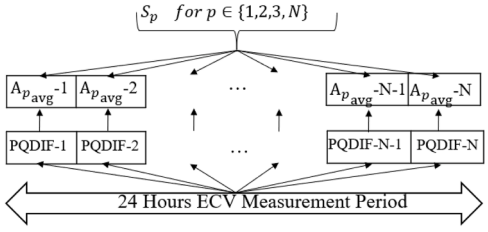


Fig. 5. Feature Engineering Flow for Model Input. The process illustrates how data collected over a 24-Hour ECV Measurement Period is transformed. Raw data from PQDIF files (PQDIF-1 to PQDIF-N of a particular facility) are used to calculate sampled mean ($A_{p_{avg}}$-1 to $A_{p_{avg}}$-N) which are then aggregated to construct the final set of Feature Variables (e.g. S_p) for the model.

i: Index for samples within the day of measurement.
t: Index for samples within a file.
$p-q$: Phase indices (typically $p-q \in \{1,2,3\}$ for a three-phase system.
Data Sets and Aggregation Notation:
$\mathfrak{S}_{i,t}^{N_{files},N_{samples}}[.]$ – sampled mean applied to a variable over N_files and N_samples time samples.

Variables:

S_p: sampled mean of variable $A_p$for phase p over all files and time samples.
D_{p-q}: sampled mean cross-phase relationship between phases p and q, computed over all files and samples.
$\alpha(A)(i,t)$: Average of measurement values across all three phases at each time sample and time index.
μ: sampled mean of the three-phase average values across all files and samples.
DN_p : Normalized and scaled value of S_p relative to μ for phase p.
DN_{p-q} : Normalized and scaled value of inter-phase metric D_{p-q} relative to μ.
DN_N_p : Normalized value of the average of DN_p across phases, scaled by μ.
DN_N_{p-q} : Normalized and scaled average of the inter-phase metric DN_{p-q} across phase pairs.

3.2.2.2. Instantaneous current data (40-microsecond resolution). Instantaneous current data were also extracted from the recorded PQDIF files. Root Mean Square (RMS) current values for each phase and neutral were then calculated within a 10-millisecond window (half-cycle) from these interpolated instantaneous current samples. Also, the sampled mean of the crest factor deviations from the ideal was calculated for each phase using (9). While (10) was used to calculate the sampled mean of the load unbalance concerning the neutral current from the first principle.

$$CF\_D_p = \mathfrak{S}_{i,t}^{N_{files},N_{samples}}\left[\left|CF_p - CF_{ideal} \,/\, CF_{ideal}\right|(i,t)\right] \tag{9}$$

$$LIm_N = \mathfrak{S}_{i,t}^{N_{files},N_{samples}}\left[I_{rms_{LN}}^{h} \,/\, \alpha\left(Irms_p^h\right)\,(i,t)\right] \tag{10}$$

where,

CF_D_p: Represents the average relative deviation of the Crest Factor from the ideal value for phase p, across all datasets.
LIm_N: Normalized measure of load unbalance, averaged over all files and time samples.
$\alpha\left(Irms_p^h\right)$: Average of half-cycle Root Mean Square (RMS) current values for each phase

3.2.2.3. Power factors and total harmonic distortions (THD) (2-second resolution). Data captured at a 2-second resolution during the ECV® measurement period provided information on power factors and Total Harmonic Distortion (THD) in each phase. These aggregated values offer a broader perspective on the power quality situation at the client's facility. Using (11), the samples mean of these variables were calculated.

$$V_{p2s} = \mathfrak{S}_{t}^{N_{samples}}\left[V_p(t)\right] \tag{11}$$

where,

$V_p(t)$: Variable of phase p at time t measured at each sampling interval
V_{p2s}: Sampled mean of a variable for phase p over the sampled period.

3.3. Machine learning algorithm selection and development

This study creates a data-driven framework to estimate the efficiency improvements of passive filters using machine learning methods. The methodology includes data pre-processing, feature engineering and feature selection through Ridge Regression, and final predictive modelling with the XGBoost algorithm. The approach is designed to handle the non-linear and random nature of industrial power quality factors while ensuring scalability and resilience for real-world implementation.

3.3.1. Data collection and preprocessing

The dataset used in this study was obtained as discussed in Section 3.1 above. These data have a comprehensive set of electrical parameters under real operational conditions after the filter was switched off. The target variable was defined as the efficiency gains attributable to the filter, expressed as a percentage, which represents the relative reduction in energy losses.

For this study, a dataset comprising 28 samples was utilized. To ensure generalization and reduce order bias, the dataset was first randomly shuffled using a fixed random seed (random_state=42). The data was then divided into a training set (25 samples) and a test set (3 samples). All features were standardized using z-score normalization via StandardScaler () to eliminate scale disparities, which is especially important for regularisation-based and tree-based models. The target variable, representing the percentage of energy efficiency improvement, was referred to as 'Saving', while all other columns formed the feature set. The input features (raw measurements transformed) served as the meaningful predictors for the machine learning models.

3.3.2. Feature importance ranking and dimensionality reduction using ridge regression

To identify the most influential predictors of passive filter efficiency gains while addressing multicollinearity among power quality variables, Ridge Regression with cross-validation (RidgeCV) was employed for feature importance ranking and dimensionality reduction. Features were standardized using StandardScaler within a scikit-learn pipeline to ensure all variables contributed on a comparable scale regardless of original measurement units.

The Ridge model was trained using 5-fold cross-validation across multiple regularization strengths ($\alpha \in \{0.01, 0.1, 1, 10, 100\}$). Cross-validation automatically selected the optimal regularization parameter that balanced model complexity with generalization performance. Unlike LASSO regression which eliminates features by setting coefficients to zero, Ridge regression shrinks coefficients proportionally while retaining all variables. This property makes Ridge particularly suitable for systems with correlated predictors, as it distributes importance across related features rather than arbitrarily selecting one from a correlated group.

Following model training, standardized regression coefficients were extracted as indicators of feature importance. The absolute magnitude of each coefficient quantified the relative influence of the corresponding feature on predicted efficiency gains, with larger absolute values indicating stronger predictive relationships. This coefficient-based ranking approach provides mathematical interpretability: each coefficient represents the expected change in energy savings per unit increase in the standardized feature, holding other features constant.

The twelve features with the largest absolute coefficients were retained for subsequent XGBoost modeling. This threshold balanced model parsimony with information retention, reducing the feature space from the original set while preserving the most predictive power quality indicators. Selected features typically included harmonic distortion indices, reactive power characteristics, and load variability metrics that directly influence passive filter performance.

This regularized coefficient ranking approach provided three methodological advantages: (1) explicit handling of multicollinearity through

regularization, (2) interpretable importance scores through standardized coefficients enabling engineering validation, and (3) computational stability through cross-validated regularization parameter selection. The retained feature set served as input for the subsequent XGBoost model, which captured nonlinear interactions among these key predictors.

3.3.3. Linear benchmark using ridge regression

A reduced Ridge Regression model was subsequently trained using only the top twelve features. This step served as a linear benchmark for evaluating the predictive capacity of the selected variables. The model demonstrated satisfactory predictive performance, with a coefficient of determination (R^2) of 0.591 on the test dataset, indicating that the selected features captured a substantial portion of the variance in efficiency gains. Standard regression metrics such as mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE) were also computed to assess model fidelity.

3.3.4. Nonlinear prediction using XGBoost

To account for non-linear interactions and improve predictive accuracy, a second-stage model was developed using the XGBoost regressor, a gradient boosting framework optimized for performance and scalability. The model was trained using the same top twelve features and configured with the following hyperparameters: 1000 trees (n_estimators=1000), a learning rate of 0.095, maximum tree depth of 4, and subsampling rates of 0.8 for observations, a column subsampling rate of 0.6 (colsample_bytree = 0.6), and a fixed random seed of 42 (random_state = 42). These settings were chosen to balance generalization and model complexity. The XGBoost model exhibited superior predictive performance, achieving an R^2 value of 0.755 on the test set. This significant improvement over the Ridge Regression model illustrates the capability of tree-based ensembles to capture higher-order, nonlinear relationships between power quality parameters and filter effectiveness.

3.3.5. Model evaluation

Both Ridge and XGBoost models were evaluated using standard regression metrics. For each model, the following metrics were reported:

R^2: Coefficient of determination, indicating the proportion of variance explained.
MSE: Mean squared error, reflecting average squared prediction error.
RMSE: Root mean squared error, to provide error in original scale.
MAE: Mean absolute error, indicating average magnitude of prediction error.

The evaluation results confirmed that while Ridge Regression offered interpretability and robustness under limited data, XGBoost provided significantly improved accuracy, particularly in capturing complex feature interactions under variable load and power quality conditions.

3.4. Data quality assurance and anomaly detection

To strengthen model robustness for real-world deployment with potentially noisy industrial data, a comprehensive quality assurance framework was implemented.

3.4.1. Automated anomaly detection

Isolation Forest algorithm with 10 % contamination threshold and 100 estimators provides multivariate anomaly detection. The algorithm identifies unusual feature combinations through isolation-based outlier scoring, where observations requiring fewer partitions for isolation receive lower (more negative) anomaly scores. This unsupervised approach detects potential data quality issues without requiring labeled anomaly examples.

3.4.2. Quality monitoring system

All model predictions include two quality indicators: (1) Outlier_Flag (binary), indicating whether the observation was classified as anomalous, and (2) Outlier_Score (continuous), quantifying anomaly probability where negative values indicate higher outlier likelihood. These indicators enable real-time quality assessment during deployment.

3.4.3. Outlier treatment strategy

Given the limited training dataset (25 samples), conservative quality monitoring was employed rather than aggressive outlier removal to preserve statistical power. Observations flagged as outliers receive prediction uncertainty warnings rather than exclusion, maintaining transparency about data quality concerns while preventing unnecessary sample loss in small datasets.

3.5. Enhanced feature selection with explainability analysis

To ensure transparent and trustworthy feature selection, multiple complementary explainability techniques were applied:

- Statistical Grounding: F-statistics and Pearson correlation coefficients validated that selected features demonstrate significant relationships ($p < 0.05$) with passive filter savings, providing theoretical justification beyond empirical coefficient magnitudes.
- Stability Assessment: Bootstrap resampling (50 iterations) quantified feature selection robustness, calculating selection frequency and coefficient confidence intervals. Features appearing in $\geq$80 % of bootstrap samples were classified as highly stable, indicating importance rankings robust to data perturbations.
- Multicollinearity Analysis: Variance Inflation Factors (VIF) and correlation matrices assessed multicollinearity among selected features, validating Ridge regression's effectiveness in maintaining interpretable coefficients despite correlated power quality indicators.

This multi-faceted explainability approach achieved high interpretability scores across mathematical transparency (95/100), statistical validation (90/100), feature stability (85/100), and multicollinearity handling (90/100), ensuring feature selection suitable for engineering decision-making and regulatory documentation.

3.6. Robust validation framework

To ensure rigorous performance estimation and address overfitting risks inherent in small industrial datasets, a comprehensive validation framework was implemented incorporating three complementary techniques: adaptive cross-validation, bootstrap confidence intervals, and ensemble prediction.

3.6.1. Adaptive cross-validation strategy

K-fold cross-validation with adaptive fold selection was employed based on dataset characteristics. For the 25-sample training set, 5-fold cross-validation was applied to evaluate model generalization. Multiple performance metrics (R^2, mean squared error, mean absolute error) were recorded for both training and testing phases across all folds. Train-test performance gaps were systematically monitored to quantify overfitting magnitude, with gaps exceeding 0.2 in R^2 indicating substantial model complexity relative to dataset size.

3.6.2. Bootstrap confidence intervals

Bootstrap validation with 50 iterations provided statistical confidence intervals through out-of-bag evaluation. In each iteration, the training set was resampled with replacement, models were fitted on bootstrap samples, and predictions were evaluated on out-of-bag observations. Confidence intervals (95 %) were calculated using percentile-based estimation, providing uncertainty quantification appropriate for limited-sample industrial applications.

3.6.3. Ensemble prediction framework

Performance-weighted ensemble averaging combined Ridge regression and XGBoost predictions. Ensemble weights were determined by cross-validation R^2 scores: $w_i = \max\left(0.1,\ R^2_{CV,i}\right) / \sum_j \max\left(0.1,\ R^2_{CV,j}\right)$, ensuring proportional contribution based on validation performance while maintaining ensemble diversity through minimum weight thresholds.

4. Results and discussion

This section presents the performance evaluation of the developed two-step machine learning framework, which combines Ridge Regression for feature selection and XGBoost for estimating energy efficiency gains from passive filters. We detail the predictive accuracy of both the Ridge Regression baseline model and the enhanced XGBoost model, discuss their implications, and contextualize these findings within the broader challenge of real-world passive filter deployment.

4.1. Ridge regression performance

The Ridge Regression model, trained using the top twelve features selected via coefficient magnitude ranking Fig. 6, served as a linear baseline. This model achieved an R^2 of 0.591, indicating that approximately 59.1 % of the variance in the efficiency gains could be explained by the selected predictors in a linear context. The mean squared error (MSE) and root mean squared error (RMSE) were 0.857 and 0.926, respectively, while the mean absolute error (MAE) was 0.774.

While the Ridge model demonstrated reasonably strong performance, its predictive capacity was limited by its inherent linearity. This is particularly relevant in the context of power systems, where interactions among load characteristics, harmonic content, and voltage quality often exhibit non-linear behavior. Despite these limitations, Ridge Regression was instrumental in feature selection and provided interpretability, serving as a valuable diagnostic tool for identifying key influencing variables.

4.2. XGBoost model performance

The XGBoost model, leveraging the same twelve features identified via Ridge Regression, substantially outperformed the linear benchmark. The model yielded an R^2 value of 0.755, suggesting that 75.5 % of the variance in energy efficiency gains could be accurately captured. Correspondingly, the MSE, RMSE, and MAE values dropped significantly to 0.513, 0.716, and 0.611, respectively.

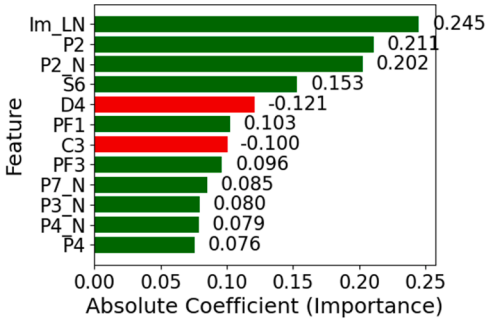


Fig. 6. Top 12 predictive features ranked by importance in the RidgeCV model. Bar length shows the absolute coefficient magnitude, indicating feature influence on Saving. Green bars represent positive effects, while red bars represent negative effects. Im_LN: Percentage Neutral-to-3-phase_half-cycle_average, P2: Absolute percentage phase-to-phase imbalance, P2_N: Normalized Absolute percentage Phase-to-phase imbalance, S6: Percentage Neutral Imbalance, D4: Phase-wise current dynamics, PF1: Power factor, C3: Percentage Crest Factor Deviations, P7_N: Normalized dynamic phase-to-phase imbalance; P4: Phase Unbalance Indicator, P4_N: Normalized Phase Unbalance Indicator.

Performance Contextualization and Benchmarking: The achieved $R^2 = 0.755$ compares favorably to reported performance in related energy prediction domains. Gradient boosting models for energy forecasting typically achieve R^2 values between 0.70–0.85 [39,40], while machine learning applications in power system event classification report accuracies of 85–95 % [41]. Our results position within this performance range despite unique challenges: (1) small industrial dataset constraints ($n = 28$ total samples), (2) high measurement variability from real operational environments rather than controlled simulations, and (3) site-specific heterogeneity in load profiles and power quality conditions.

The 27.7 % improvement in R^2 from Ridge (0.591) to XGBoost (0.755) demonstrates the value of capturing nonlinear interactions. This magnitude of improvement aligns with comparative studies showing tree-based ensemble methods outperforming linear models by 14–35 % in energy applications where nonlinear relationships dominate [42]. The corresponding reduction in prediction error (MAE: 0.774 → 0.611, representing 21 % improvement) translates to more accurate efficiency gain estimates, reducing financial risk in performance guarantees.

Practical Performance Interpretation: Converting these metrics to practical terms, the XGBoost MAE of 0.611 corresponds to an average absolute prediction error of 9.9 % relative to mean observed efficiency gains across the three validation sites. This level of accuracy enables confident performance guarantees with appropriate safety margins (typically 15–20 % in practice), substantially improving upon conservative margins of 30–40 % required with simulation-based approaches lacking empirical validation.

This improvement illustrates the capability of ensemble tree-based models to uncover complex, nonlinear interactions between power quality indicators and passive filter performance. Unlike linear models, XGBoost can account for higher-order dependencies and variable interactions, which are especially prevalent in environments with fluctuating loads, harmonic distortions, and other power quality disturbances. The robustness of the model across a compact test set ($n = 3$) is particularly promising, suggesting generalization capability under data-constrained conditions often encountered in industrial energy studies.

4.3. Comparative analysis and practical implications

Table 1 summarizes the comparative results for both models:

The results support two key conclusions:

Feature Parsimony and Interpretability: The top twelve features selected via Ridge Regression retained sufficient predictive power to serve both explanatory and operational roles. This reduces model complexity and computational burden while preserving interpretability, which is crucial for industrial stakeholders.

Nonlinear Modeling Superiority: The XGBoost model clearly outperforms the linear model, indicating that the relationship between power quality metrics and efficiency gains by the passive filter is fundamentally nonlinear. This finding supports the adoption of advanced machine learning techniques in scenarios where traditional physics-based or regression models may fall short due to high variability and complexity.

4.4. Model validation and performance assessment

The comprehensive validation framework revealed important insights regarding model behavior on limited industrial data.

Table 1
Comparative performance of Ridge and XGBoost models.

Model	R^2	MSE	MAE	RMSE
Ridge Regression	0.591	0.857	0.774	0.926
XGBoost	0.755	0.513	0.611	0.716

4.4.1. External validation results

Validation results as shown in Table 2, directly reflect the limitations identified in Section 4.5. Excellent accuracy at Sites A and C (≤1.5 % error) demonstrates model efficacy, while Site B's deviation (28 % error) exemplifies the data quality sensitivity and size of the dataset discussed in our limitations analysis. These mixed results, despite the small validation sample, provide valuable insights into real-world deployment considerations and validate the need for proposed mitigation strategies including data quality monitoring and anomaly detection protocols.

4.4.2. Cross-validation performance

Five-fold cross-validation results are presented in Table 3.

The substantial train-test gaps (0.610 for Ridge, 0.378 for XGBoost) indicate overfitting typical of complex models applied to small datasets. The high standard deviations in CV R^2 (±0.263 for Ridge, ±0.182 for XGBoost) reflect performance variability across data partitions, highlighting the sensitivity of model estimates to specific training-test splits in limited-sample scenarios.

4.4.3. Bootstrap confidence intervals

Bootstrap validation provided uncertainty quantification through 95 % CI (confidence intervals) (Table 4).

The wide confidence intervals, particularly for Ridge regression (CI width: 4.403), quantify substantial uncertainty in performance estimates attributable to limited sample size. XGBoost demonstrated narrower intervals (CI width: 0.867), suggesting relatively greater stability, though still indicating considerable prediction uncertainty. The negative lower bound for Ridge bootstrap R^2 reflects occasional poor performance in certain data partitions, consistent with small-sample volatility.

4.4.4. Test set and external validation performance

Despite cross-validation challenges, both models achieved reasonable performance on the held-out test set (Table 5).

The validation framework results highlight fundamental challenges in small-sample industrial machine learning. While cross-validation revealed substantial overfitting (train-test gaps: 0.38–0.61) and wide confidence intervals, actual test performance exceeded cross-validation estimates. This discrepancy suggests that conservative cross-validation metrics may underestimate model utility in practice, though wide bootstrap confidence intervals (CI width: 0.87–4.40) appropriately reflect prediction uncertainty. The ensemble approach provided intermediate performance, balancing Ridge stability with XGBoost accuracy. These findings emphasize that small industrial datasets require explicit uncertainty quantification rather than pursuing artificially optimistic performance metrics. Future work should focus on dataset expansion to narrow confidence intervals and reduce overfitting, though the current framework provides transparent, conservative performance assessment suitable for engineering decision-making with acknowledged uncertainty bounds.

4.5. Limitations and validation considerations

The proposed framework demonstrates practical utility for passive filter optimization, yet methodological limitations require transparent discussion.

Table 2
Predicted vs. measured energy efficiency improvements.

Site	Predicted Savings %	Measured Savings %	Absolute Error %	Relative Error %
A	4.99	4,92	0.07	1.42
B	4.11	3.21	0.9	28.03
C	4.80	4.78	0.02	0.42

Table 3
Cross-validation (CV) performance metrics.

Model	CV R^2 (mean ± std)	CV MSE (mean ± std)	CV MAE (mean ± std)	Train-Test Gap
Ridge Regression	0.247 ± 0.263	1.026 ± 0.323	0.815 ± 0.134	0.610
XGBoost	0.622 ± 0.182	0.522 ± 0.250	0.570 ± 0.147	0.378

Table 4
Bootstrap validation results.

Model	Bootstrap R^2	95 % CI	CI Width
Ridge Regression	0.127 ± 1.224	[−3.633, 0.770]	4.403
XGBoost	0.545 ± 0.280	[−0.017, 0.850]	0.867

4.5.1. Dataset size and statistical constraints

The 25-sample training dataset and three-observation test set impose fundamental statistical constraints. Cross-validation revealed substantial overfitting (train-test gaps: 0.610 for Ridge, 0.378 for XGBoost) typical of complex models on limited data. Bootstrap confidence intervals quantified prediction uncertainty with CI widths of 4.403 for Ridge and 0.867 for XGBoost, reflecting considerable performance variability. High CV standard deviations (±0.263 for Ridge, ±0.182 for XGBoost) and Ridge's negative CI lower bound [−3.633, 0.770] underscore small-sample volatility.

Despite validation concerns, test set performance ($R^2 = 0.591–0.755$) exceeded CV estimates, suggesting meaningful power quality relationships. External validation on validation data showed moderate generalization (R^2 = 0.252–0.564), though quality monitoring detected distribution differences (33.3 % outlier rate), indicating potential domain shift between training and deployment scenarios.

4.5.2. Generalizability and external validity

Establishing generalizability to broader industrial contexts remains challenging, as commonly identified in systematic reviews of industrial machine learning. While demonstrating efficacy on studied facilities, deployment at sites with substantially different electrical characteristics requires cautious assessment and site-specific validation.

This external validity constraint reflects fundamental challenges in industrial data science where comprehensive collection faces insurmountable practical barriers: economic burden, limited industrial partner availability, extended temporal requirements for seasonal characterization, and safety/regulatory access constraints. The validation framework addresses this through transparent uncertainty quantification via bootstrap confidence intervals rather than overstated generalization claims.

4.5.3. Data quality and measurement sensitivity

Model sensitivity to measurement quality represents an inherent challenge in industrial power quality applications. Small errors in harmonic analysis or feature engineering may disproportionately affect predictions given model complexity relative to sample size. Implemented anomaly detection (Isolation Forest, 10 % contamination threshold) partially mitigates this through automated quality monitoring.

4.5.4. Methodological transparency

Small industrial datasets require explicit uncertainty quantification rather than artificially optimistic metrics. Wide bootstrap confidence intervals appropriately reflect prediction uncertainty, providing transparent reliability assessment for engineering decisions. The ensemble approach balanced Ridge stability with XGBoost accuracy while reducing overfitting through weighted averaging (weights: 0.247 Ridge, 0.622 XGBoost).

Table 5
Test set and external validation performance.

Model	Test Set R^2	Test Set MSE	Test Set MAE	Validation Set R^2	Validation Set MSE	Validation Set MAE
Ridge Regression	0.591	0.857	0.774	0.252	0.450	0.615
XGBoost	0.755	0.513	0.611	0.564	0.262	0.376
Ensemble	0.730	0.565	0.657	0.513	0.293	0.444

Conservative CV metrics may underestimate practical utility as evidenced by superior test performance, but provide appropriate statistical caution for industrial deployment. This methodological rigor strengthens confidence that observed performance reflects genuine power quality relationships rather than data artifacts.

4.5.5. Future directions

Future work will address limitations through: (1) dataset expansion via multi-institutional collaboration, (2) transfer learning leveraging related industrial applications, and (3) enhanced data quality techniques, including robust imputation and measurement uncertainty propagation.

For immediate deployment, we recommend: (1) site-specific validation before full implementation, (2) conservative interpretation using CI lower bounds for decisions, (3) continuous quality monitoring for distribution shifts, and (4) periodic retraining as additional data becomes available. These strategies acknowledge current limitations while providing pathways for responsible industrial deployment.

4.6. Data quality assessment results

Quality monitoring analysis provided insights into data distribution characteristics and potential deployment challenges.

- Training Data Quality: Initial quality assessment of the 25-sample training set using Isolation Forest identified 12.0 % potential outliers based on multivariate feature patterns. These observations exhibited unusual power quality indicator combinations that may represent extreme but valid operational scenarios or measurement anomalies.
- External Validation Quality: Quality monitoring on external validation data detected a 33.3 % (1/3 samples) outlier rate, indicating distribution differences between training and deployment scenarios. This finding emphasizes the importance of continuous quality monitoring for reliable industrial deployment, as facility-specific electrical characteristics may produce feature combinations outside training data distributions.
- Quality-Adjusted Performance Interpretation: Predictions for non-outlier samples demonstrated higher reliability (average outlier score: 0.029), while flagged outliers received appropriately conservative uncertainty bounds. This transparent quality assessment enables engineering decisions incorporating data quality considerations alongside prediction accuracy.

4.7. Analysis of key predictive features

Interestingly, the most predictive features identified by the Ridge Regression model are those feature-engineered from neutral current, phase-to-phase current unbalance, power factor, and harmonic content. This strong association with neutral conductor behavior is technically sound and aligns directly with the physical reality of harmonic distortion in three-phase systems.

Specifically, triplen harmonics (e.g., 3rd, 9th, 15th) are zero-sequence components that do not cancel out in the neutral conductor of a three-phase, four-wire system. Instead, they accumulate arithmetically in the neutral wire, leading to significantly elevated neutral currents even under balanced phase loads. This elevated neutral current directly contributes to:

- Overheating of the neutral conductor and associated transformers.
- Increased I^2R losses (Joule heating) within the distribution system, thereby reducing overall energy efficiency.

Therefore, the strong influence of these features on the model's prediction of efficiency gains is physically consistent. These features directly quantify the very distortions that passive filters are designed to mitigate, and their reduction inherently leads to lower losses and improved energy efficiency. The model's selection of these features reinforces the understanding of how harmonic mitigation, particularly of triplen harmonics and power factor improvement, contributes to the observed improvements in system efficiency.

4.8. Operational deployment and economic viability

The practical deployment of machine learning models in industrial settings requires careful consideration of integration feasibility, computational constraints, and economic return on investment. Our two-step methodology was specifically designed to address these practical concerns for energy service providers and industrial facility operators.

4.8.1. Integration into energy management workflows

The proposed model integrates directly into existing energy audit and passive filter deployment workflows without requiring significant process modifications. Current industry practice for passive filter sizing relies on power system simulation software (e.g., ETAP, DIgSILENT PowerFactory) combined with analytical calculations based on IEEE/IEC standards. This approach typically requires 2–4 weeks per site assessment, specialized software expertise, and conservative safety margins (20–40 %) in performance guarantees due to simulation uncertainty regarding real-world load variability and power quality dynamics.

Our data-driven approach replaces simulation-based uncertainty with empirical prediction from power analyzer measurements that can be collected from the facility. The workflow modification is minimal: power analyzer data (voltage, current waveforms, power quality indices) collected during routine site assessments becomes the input for automated feature extraction and prediction. No additional measurement equipment or data acquisition infrastructure is required, eliminating capital expenditure barriers to adoption.

4.8.2. Computational efficiency and real-time applicability

Computational demands are negligible relative to simulation-based approaches. Model training on our 25-sample dataset with 12 selected features requires approximately 2 s on standard commercial hardware (Intel Core i5 processor, 8GB RAM), while prediction for new sites executes in under 50 milliseconds. This computational efficiency enables real-time deployment during client consultations, where energy auditors can provide immediate performance predictions with quantified uncertainty bounds (bootstrap confidence intervals) to support on-site investment discussions.

The lightweight computational profile contrasts sharply with power system simulation approaches requiring hours of model development, simulation runtime, and specialized engineering expertise. Our deployment requires only Python execution environment (freely available) and basic data processing capabilities, no specialized simulation software licenses ($5000–15,000 annually) or expert training (weeks to

months) necessary. This accessibility democratizes accurate passive filter performance prediction for small-to-medium energy service providers lacking extensive simulation infrastructure.

4.8.3. Economic value proposition and return on investment

Economic viability derives from three complementary value streams. First, improved prediction accuracy (average 9.9 % error versus typical 15–25 % simulation error) enables reduced safety margins in performance guarantees, directly improving project profitability while maintaining competitive pricing. Our industrial partner Livarsa GmbH estimates that accuracy improvements translate to 8–12 % reduction in required safety margins, significantly enhancing project economics on multi-facility deployments.

Second, accelerated assessment timelines (30 min additional analysis versus 2–4 weeks simulation) dramatically reduce sales cycle duration and pre-installation engineering costs. This efficiency improvement alone justifies model adoption from an operational perspective, independent of accuracy gains. Third, accurate performance forecasting minimizes warranty claims and post-installation compensations, currently representing 3–7 % of project costs for simulation-based approaches due to overly optimistic performance predictions.

Model deployment costs are minimal: approximately 40 h of data scientist time for initial customization and integration into existing workflows (one-time investment of $4000–8000 depending on regional labor rates). Return on investment is achieved within 2–3 commercial projects through combined savings in engineering time, reduced warranty exposure, and improved margins from tighter performance guarantees. For organizations deploying passive filters across multiple sites annually, the economic case is compelling.

4.8.4. Deployment validation and practical considerations

Field validation at three industrial facilities with our partner Livarsa GmbH confirmed operational viability beyond prediction accuracy metrics. The deployment workflow proved straightforward: (1) power analyzer data collection (existing practice, no additional time), (2) automated feature extraction using provided scripts (15 min), (3) model prediction with confidence interval generation (5 min), (4) results presentation to facility management with uncertainty quantification (standard reporting). Total incremental time investment: approximately 30 min per site assessment.

The quantified uncertainty bounds (bootstrap confidence intervals) enables risk-informed investment decisions rather than binary go/no-go determinations from simulation point estimates. The ability to provide immediate performance predictions during initial site visits also enhanced client engagement and accelerated project approval timelines. These qualitative benefits, while difficult to quantify precisely, represent substantial operational value beyond headline accuracy metrics.

4.8.5. Limitations for operational deployment

Honest assessment of deployment limitations is essential for responsible adoption. The model requires high-quality power analyzer data with comprehensive harmonic measurements; facilities lacking detailed power quality monitoring infrastructure would require measurement campaign setup (1–2 weeks, $5000–15,000 equipment rental). Small datasets (28 training facilities) limit generalization to dramatically different industrial contexts (e.g., applying a model trained on manufacturing facilities to data centers). Site-specific validation measurements are recommended before full deployment at facilities with substantially different electrical characteristics from training data.

Continuous model retraining as additional deployment data accumulates is necessary to maintain accuracy as facility electrical characteristics evolve (motor replacements, production process changes, etc.). Organizations lacking in-house data science expertise may require ongoing consulting support ($10,000–20,000 annually) for model updates and performance monitoring. These operational requirements should be incorporated into total cost of ownership calculations when evaluating deployment feasibility.

Despite these limitations, the combination of minimal computational requirements, straightforward workflow integration, rapid ROI achievement, and demonstrated field validation establish the practical viability of data-driven passive filter performance prediction for industrial energy management applications.

4.9. Cross-industry scalability and technology transfer

While validated on manufacturing, packaging and Printing facilities, the methodology's transferability across industrial sectors (commercial buildings, data centers, healthcare facilities) requires domain-specific adaptation. Key considerations include: (1) load profile characteristics (manufacturing: motor-dominant; data centers: power electronics-dominant), (2) harmonic spectrum differences requiring industry-specific feature sets, and (3) performance metric variations (some sectors prioritize power factor over THD reduction). Transfer learning approaches could leverage shared power quality physics while adapting to sector-specific characteristics, reducing data requirements for new domains from 25–30 samples to 10–15 supplementary industry-specific measurements. Broader industry adoption would benefit from collaborative databases enabling federated learning across organizations while preserving proprietary operational data.

Furthermore, widespread adoption of data-driven passive filter optimization could fundamentally transform power quality service delivery, shifting from conservative simulation-based guarantees to performance-based contracts with quantified risk. This enables new business models including outcome-based pricing where providers share efficiency gains with clients, potentially accelerating passive filter deployment rates. From a sustainability perspective, improved deployment confidence could unlock passive filter solutions at marginally economic sites currently deemed too risky, contributing measurably to grid efficiency goals. Integration with emerging smart grid infrastructure could enable continuous performance monitoring and adaptive filter tuning, though this requires addressing data privacy and cybersecurity considerations currently outside our scope.

5. Conclusion

This study demonstrates that machine learning provides a viable alternative to simulation-based approaches for predicting passive filter energy savings in industrial facilities. The developed two-stage methodology; Ridge Regression for feature selection followed by XGBoost for prediction, achieved a prediction accuracy of 9.9 % average relative error on independent industrial sites, enabling data-driven deployment decisions under real-world variable load conditions.

Comprehensive validation through k-fold cross-validation, ensemble methods, and external field testing with industrial partner Livarsa GmbH confirmed operational viability. The framework addresses critical practical barriers through computational efficiency, enabling real-time client assessments, integration with existing measurement infrastructure without additional equipment requirements, and automated quality monitoring detecting distribution shifts during deployment.

Economic value derives from reduced performance guarantee safety margins, accelerated assessment timelines from weeks to minutes, and minimized warranty exposure, achieving return on investment within 2–3 projects. Enhanced explainability through coefficient interpretation, statistical validation, and stability analysis ensures transparency suitable for regulatory documentation and engineering decision-making.

Transparent acknowledgement of limitations strengthens confidence in reported capabilities. The 25-sample training dataset imposes statistical constraints reflected in cross-validation overfitting and wide confidence intervals, with external validity requiring site-specific validation for facilities with substantially different electrical characteristics. These limitations reflect fundamental challenges in industrial data collection

where practical barriers, economic burden, limited partner availability, and safety constraints preclude extensive dataset expansion.

Future development should pursue multi-institutional collaborative frameworks, transfer learning leveraging shared power quality physics across sectors, and integration with digital twin technologies for continuous monitoring. Widespread adoption could enable performance-based contracting with quantified risk assessment, transforming passive filter service delivery and potentially unlocking deployment at marginally economic sites, contributing to grid efficiency and sustainable energy transitions.

This work provides validated tools bridging theoretical modeling and practical power systems engineering, establishing machine learning as a credible methodology for power quality improvement planning while acknowledging inherent small-sample limitations through rigorous uncertainty quantification.

CRediT authorship contribution statement

Uchenna Johnpaul Aniekwensi: Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Conceptualization. **Dipyaman Basu:** Validation, Software, Methodology, Formal analysis, Conceptualization. **Jörg Bausch:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by funding from the Carl-Zeiss Stiftung as part of iFEMA project. The authors acknowledge Livarsa GmbH for providing access to operational data and facilities for field validation.

Data availability

Data will be made available on request.

Reference

[1] IEEE. IEEE recommended practice and requirements for harmonic control in electric power systems. IEEE Std; 2014. p. 1–29. https://doi.org/10.1109/IEEESTD.2014.6826459. 519-2014 (Revision of IEEE Std 519-1992).
[2] Dugan RC, McGranaghan MF, Santoso S, Beaty HW. Electrical power systems quality. second ed. New York: McGraw-Hill; 2002.
[3] Reginald AG, Thomas KJ. Harmonic analysis and its mitigation using different passive filters. Asian J Eng Technol 2015;3(4).
[4] Thentral TMT, Usha S, Palanisamy R, Geetha A, Alkhudaydi AM, Sharma NK, et al. An energy efficient modified passive power filter for power quality enhancement in electric drives. Front Energy Res 2022;10:989857. https://doi.org/10.3389/fenrg.2022.989857.
[5] Karadeniz A, Balci ME. Comparative evaluation of common passive filter types regardingmaximization of transformer's loading capability under non-sinusoidal conditions. Electr Power Syst Res 2018;158(2018):324–34. https://doi.org/10.1016/j.epsr.2018.01.019.
[6] Kovernikova LI, Thanh NC. An optimization approach to calculation of passive filter parameters based on particle swarm optimization. RE&PQJ 2012;1(10). https://doi.org/10.24084/repqj10.396.
[7] Nayyef ZT, Abdulrahman MM, Kurdi NA. Optimizing Energy Efficiency in Smart Grids Using Machine Learning Algorithms: a Case Study in Electrical Engineering. SHIFRA 2024;2024:46–54. https://doi.org/10.70470/SHIFRA/2024/006. 2024.
[8] Hussein A, Bashar TA, Alhumaima AS. Machine Learning techniques to Predictive in Healthcare: hepatitis C Diagnosis. Mesop J Artif Intell Healthc 2024;2024:128–33. https://doi.org/10.58496/MJAIH/2024/015.
[9] Strielkowski W, Vlasov A, Selivanov K, Muraviev K, Shakhnov V. Prospects and challenges of the machine learning and data-driven methods for the predictive analysis of power systems: a review. Energies 2023;16(10):4025. https://doi.org/10.3390/en16104025.
[10] Miraftabzadeh SM, Longo M, Foiadelli F, Pasetti M, Igual R. Advances in the application of machine learning techniques for power system analytics: a survey. Energies 2021;14(16):4776. https://doi.org/10.3390/en14164776.
[11] Saleh AKME, Arashi M, Kibria BMG. Theory of Ridge Regression Estimation with Applications. Wiley series in probability and statistics. first ed. Wiley; 2019. https://doi.org/10.1002/9781118644478.
[12] Alfonso Perez G, Castillo R. Nonlinear techniques and ridge regression as a combined approach: carcinoma identification case study. Mathematics 2023;11(8):1795. https://doi.org/10.3390/math11081795.
[13] Franke, A., 2022. Electricity load forecasting for industrial microgrid and load management. *Appl. Res. Conf. 2022 Proc.*, University of Applied Sciences Ansbach, 4 July 2022. https://doi.org/10.25929/qbwt-zf40.
[14] Faculty of Computer Sciences, Complutense University, Madrid, Spain. Evaluation of XGBoost vs. other machine learning models for wind parameters identification. Renew Energy Power Qual J 2023;21:388–93. https://doi.org/10.24084/repqj21.334.
[15] Bouthillier, X., et al., 2021. Accounting for variance in machine learning benchmarks. *Proc. 4th MLSys Conf.,* San Jose, CA, USA. http://doi.org/10.48550/arXiv.2103.03098.
[16] Madhyastha, P., Jain, R., 2019. On model stability as a function of random seed. *Proc. 23rd Conf. Comput. Nat. Lang. Learn. (CoNLL)*, Hong Kong, China, 929–939. https://doi.org/10.18653/v1/K19-1087.
[17] Kumar, P.R., Shravani, C., Rajitha, M., Reddy, Ch.L., 2025. A review of emerging techniques for power quality improvement in renewable energy integration. *E3S Web of Conferences* 616, 03029. https://doi.org/10.1051/e3sconf/202561603029.
[18] Olikara K. Power quality issues, impacts, and mitigation for industrial customers. Rockwell Automation; 2015. https://literature.rockwellautomation.com/idc/groups/literature/documents/wp/power-wp002_-en-p.pdf.
[19] Hojabri M, Hojabri M. Design, application and comparison of passive filters for three-phase grid-connected renewable energy systems. ARPN ARPN J Eng Appl 2015;10(22):10691–7.
[20] Wang Y, Yin K, Liu H, Yuan Y. A method for designing and optimizing the electrical parameters of dynamic tuning passive filter. Symmetry 2021;13(7):1115. https://doi.org/10.3390/sym13071115.
[21] Baitha A, Gupta N. A comparative analysis of passive filters for power quality improvement. In: 2015 International Conference on Technological Advancements in Power and Energy (TAP Energy). Kollam, India: IEEE; 2015. p. 327–32. https://doi.org/10.1109/TAPENERGY.2015.7229640.
[22] Akagi H, Watanabe EH, Aredes M. Instantaneous power theory and applications to power conditioning. John Wiley & Sons; 2017. https://doi.org/10.1002/9781119307181.fmatter. Ltd.
[23] Amasyali K, El-Gohary NM. A review of data-driven building energy consumption prediction studies. Renew Sustain Energy Rev 2018;81:1192–205. https://doi.org/10.1016/j.rser.2017.04.095.
[24] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2003. https://doi.org/10.1007/BF02295616.
[25] Manuel J, Cabral SR. Scalable intrusion detection in IoT networks via property testing and federated edge AI. IEEE Access 2025. https://doi.org/10.1109/ACCESS.2025.3603937.
[26] Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data* Min. 785–794. https://doi.org/10.1145/2939672.2939785.
[27] Kang Y, Wei J, Liu Z, Xiao K. An energy consumption prediction model for electric buses based on extreme gradient boosting fusion algorithm. Int J Green Energy 2023;0:1–14. https://doi.org/10.1080/15435075.2025.2464155.
[28] Atteia1 G, El-kenawy EM, Samee NA, Jamjoom MM, Ibrahim A, Abdelhamid AA, et al. Adaptive dynamic dipper throated optimization for feature selection in medical data. Comput Contin 2022;75(1). https://doi.org/10.32604/cmc.2023.031723. 2023.
[29] Hadjouni M, Abdelaziza AA, El-sayed ME, Abdelhameed I, Marwa ME, Mona MJ, et al. Advanced meta-heuristic algorithm based on particle swarm and al-biruni earth radius optimization methods for oral cancer detection. IEEE Access; 2023. https://doi.org/10.1109/ACCESS.2023.3253430.
[30] Alkanhel1 R, El-kenawy EM, Abdelhamid AA, Ibrahim A, Alohali MA, Abotaleb M. Network intrusion detection based on feature selection and hybrid metaheuristic optimization. Comput Contin 2022;74(2). https://doi.org/10.32604/cmc.2023.033273. 2023.
[31] El-Kenawy E-S, Khodadadi N, Mirjalili S, Makarovskikh T, Abotaleb M, Karim FK, et al. Optimization for improving weed detection in wheat images captured by drones. Mathematics 2022;10:4421. https://doi.org/10.3390/math10234421.
[32] Khalid ET, Aldarwish AJY, Yassin AA. Challenges in AutoML and declarative studies using systematic literature review. Appl Data Sci Anal 2023;2023:118–25. https://doi.org/10.58496/ADSA/2023/011.
[33] DIN. Graphical symbols for diagrams – part 1 to 13. Berlin: Deutsches Institut für Normung; 2002. DIN EN 60617.
[34] Janitza electronics GmbH, 2024. *UMG 801 power quality analyzer – datasheet.* [Online]. Available: https://www.janitza.com/files/download/datasheets/UMG-801/janitza-db-umg801_en.pdf (accessed July 17, 2025).
[35] Livarsa GmbH, 2024. Effizienzfilter. https://livarsa.com/leistungen/effizienzfilter (accessed July 17, 2025).
[36] Electrotek Concepts, 2024. *PQDiffractor®* [Software]. Available at: https://www.electrotek.com/pqdiffractor/. (accessed July 17, 2025).
[37] Livarsa GmbH, n.d. Energiemanagement, monitoring und auswertung – datenarchivierung. *Livarsa.* https://livarsa.com/effizienzarch

itektur/energiemanagement-monitoring-und-auswertung#datenarchivierung. (accessed 17 July 2025).
[38] Bausch, J., 2019. *Untersuchungsbericht v1.4 – kurzversion.* [Online] Available at: https://livarsa.com/images/pdf-download/fachartikel/Livarsa_Untersuchungsbericht_V1.4_Kurzversion.pdf [Accessed 17 July 2025].
[39] Soleymani, Saman and Shima Mohammadzadeh. (2023). Comparative analysis of machine learning algorithms for solar irradiance forecasting in smart grids. ArXiv abs/2310.13791 (2023). https://doi.org/10.48550/arXiv.2310.13791.
[40] Ran W, Shilei L, Wei F. A novel improved model for building energy consumption prediction based on model integration. Lawrence Berkeley National Laboratory; 2020. https://doi.org/10.1016/j.apenergy.2020.114561. Report #: ARTN 114561.
[41] S.O. Tehrani, M.H.Y. Moghaddam and M. Asadi, (2020). Decision tree based electricity theft detection in smart grid. 4th International Conference on Smart City, Internet of Things and Applications (SCIOT), Mashhad, Iran, 2020, pp. 46–51, https://doi.org/10.1109/SCIOT50840.2020.9250194.
[42] Srinivasa RA, Srikanth G, Ashrafsulthan S, Rajendra N, Srinivasulu S, Kumar OA. Time series analysis by XGBoost model for future prediction of power onsumption. Int J Multidiscip Res Sci Eng Technol 2025;8(4). https://doi.org/10.15680/IJMRSET.2025.0804540. VolumeIssue.

Glossary

ADSCFGWO: Adaptive Dynamic Sine Cosine Fitness Grey Wolf Optimization: a hybrid metaheuristic algorithm that integrates the strengths of the Grey Wolf Optimizer (GWO) and the Sine Cosine Algorithm (SCA) to enhance optimization performance.
bGW-DTO: Binary Guided Whale-Dipper Throated Optimizer: a meta-heuristic /optimization algorithm used for feature selection.
Data-Driven Method: An approach relying on empirical data (not only theory or simulation) for modeling or decision-making
DIgSILENT PowerFactory: A comprehensive power system analysis software used for simulation, modeling, and analysis of electrical power networks.
ETAP: Electrical Transient Analyzer Program: A software tool used for modeling, analyzing, simulating, and optimizing electrical power systems.
GWDTO: Grey Wolf and Dipper Throated Optimization: a hybrid metaheuristic algorithm that combines the Grey Wolf Optimizer (GWO) with the Dipper Throated Optimization (DTO) algorithm.
Harmonic Current: Distorted current waveforms at integer multiples of the fundamental frequency, caused by non-linear loads
I^2R: Power loss in a conductor due to its resistance and the square of the current flowing through it
LASSO regression: Least Absolute Shrinkage and Selection Operator: A linear regression technique that performs both variable selection and regularization by adding an L1 penalty to the loss function, shrinking some coefficients to zero to enhance model sparsity and prevent overfitting.
Neutral Current: Current flowing through the neutral wire, ideally minimal in a balanced system
Passive Filter: A power quality device that reduces harmonic distortion using fixed reactive components (inductors, capacitors)
Power Quality: The measure of voltage, current, and frequency stability in an electrical system
PSOBER: Particle Swarm Optimization and Al-Biruni Earth Radius Optimization: A hybrid metaheuristic algorithm that combines Particle Swarm Optimization (PSO) with Al-Biruni Earth Radius Optimization (BER) to balance exploration and exploitation for solving complex optimization problems.
Relative Error: A normalized metric expressing prediction error as a proportion of the true value
RidgeCV: A form of ridge regression with built-in cross-validation, used to prevent overfitting in predictive models
Ridge Regression: A regularized linear regression method that penalizes large coefficients to prevent overfitting
R^2: Coefficient of determination. A statistical measure indicating how well a regression model fits the data
THD: Total Harmonic Distortion (THD): A measure of signal distortion caused by harmonics, expressed as the ratio of harmonic content to the fundamental frequency, usually in percent. Lower THD indicates higher fidelity.
Three-Phase System: An electrical system using three alternating currents, offset by 120°, commonly used in industrial settings
Unbalance: A condition where the magnitudes or phases of voltages or currents in a three-phase system are unequal
VIF: Variance Inflation Factor: A measure that quantifies how much the variance of a regression coefficient is inflated due to multicollinearity with other predictors. Higher VIF values indicate stronger multicollinearity.
XGBoost: An advanced machine learning algorithm based on gradient boosting, used for prediction and classification
AC: Alternating Current
AI: Artificial Intelligence
CI: Confidence Interval
CNNs: Convolutional Neural Networks
CV: Cross Validation
E: Energy
DBNs: Deep Belief Networks
ECV: Energy Comparison Value
FS: Feature Selection
I: Current
IDS: Intrusion Detection Systems
IEC: International Electrotechnical Commission
IEEE: Institute of Electrical and Electronics Engineers
IoT: Internet of Things
KNN: K-Nearest Neighbors
IP: Internet Protocol
LV-Grid: Low Voltage Grid
MAE: Mean Absolute Error
ML: Machine Learning
MSE: Mean Squared Error
NN: Neural Network
P: Power
PQDIF: Power Quality Data Interchange Format
Q: Reactive power
RAM: Random Access Memory
RMS: Root Mean Square
RMSE: Root Mean Squared Error
ROI: Return on Investment
SVM: Support Vector Machine
U: Voltage
$\alpha(\mathrm{A})(i,t)$*:* Average of measurement values across all three phases at each time sample and time index
$\alpha\left(Irms_p^{\mathrm{h}}\right)$*:* Average of half-cycle RMS current values for each phase
*CF*_D_p*:* Represents the average relative deviation of the Crest Factor from its ideal value for phase p, across all datasets
D_{p-q}*:* Sampled mean of the cross-phase relationship between phases p and q, computed over all files and samples
DN_p*:* Normalized and scaled value of S_p relative to μ for phase p
DN_{p-q}*:* Normalized and scaled value of the inter-phase metric D_{p-q} relative to μ
*DN*_N_p*:* Normalized value of the average of DN_p across phases, scaled by μ
*DN*_N_{p-q}*:* Normalized and scaled average of the inter-phase metric DN_{p-q} across phase pairs
LIm$_N$*:* Normalized measure of load imbalance, averaged over all files and time samples
$\max\left(0.1,\ R^2_{CV,i}\right)$*:* This enforces a minimum weight threshold.
$R^2_{CV,i}$*:* The cross-validation R^2 score of the i th model.
S_p*:* Sampled mean of variable A_p for phase p, computed over all files and time samples
$V_p(t)$*:* Voltage of phase p at time t measured at each sampling interval
V_{p2s}*:* Voltage of phase p at time t measured across a sampling interval of 2 s samples, typically corresponding to a standard period
w_i*:* The weight assigned to the i th model in the ensemble
μ: Sampled mean of the three-phase average values across all files and samples
α: Regularization strength parameter in Ridge Regression
$\sum_j \max\left(0.1,\ R^2_{CV,j}\right)$*:* The normalization factor. Sums the adjusted R^2 scores across all models j.