energies

MDPI

Article

Machine Learning for Early Transformer Insulation Defect Risk Ranking

Lucas Thobejane and Bonginkosi A. Thango *

Department of Electrical and Electronic Engineering Technology, University of Johannesburg, Johannesburg 2092, South Africa; 223266000@student.uj.ac.za
* Correspondence: bonginkosit@uj.ac.za

Abstract

Early identification of transformer insulation defects during IEC 60270 induced-voltage acceptance testing could enable high-risk units to be prioritized before the guaranteed test voltage is reached. This study investigated whether early-ramp partial discharge characteristics could predict the partial discharge inception voltage (PDIV) and support defect-risk ranking. An HPWOA-optimized dual-output feedforward neural network (HPWOA-DO-FNN) was developed and compared with Linear Regression, Random Forest (RF) Regressor, Gradient Boosting Machine (GBM) Regressor, Logistic Regression Classifier, and RF Classifier. Ten features were extracted from 120 s sliding windows applied to the first 50% of IEC 60270 test records. The dataset comprised 1705 windows from 39 source files representing 18 power transformers tested in South Africa between 2019 and 2021. Five-fold file-level GroupKFold cross-validation prevented leakage between windows from the same test file. Uncertainty was estimated using 1000-replicate group-level bootstrap confidence intervals. Performance was evaluated using R^2, RMSE, area under the receiver operating characteristic curve (AUC), F1 score, and top-quartile risk capture. All models demonstrated near-zero or negative coefficients of determination (R^2) for absolute PDIV prediction (best result: GBM-Regressor PDIVRatio R^2 = 0.097, 95% bootstrap CI [−0.237, 0.354]), confirming that PDIV cannot be reconstructed from 10 early-ramp features alone—an expected result given that PDIV is a transformer-level property, while windows sample correlated sub-sequences of a single test session. Despite this, relative defect-risk ranking remains achievable: the ranking task does not require accurate point prediction. GBM-Regressor achieved the highest PDIVRatio R^2 (0.097) and lowest RMSE (0.323), although its confidence interval included zero. For defect discrimination, RF Classifier achieved an AUC of 0.714 (95% CI, 0.628–0.803) and the highest top-quartile capture (43.5%), while GBM-Regressor achieved an AUC of 0.712 (95% CI, 0.624–0.801) and an F1 score of 0.603. HPWOA achieved an AUC of 0.527 (95% CI, 0.389–0.662). Early-ramp features are insufficient for accurate transformer-level PDIV prediction but provide useful relative defect-risk ranking. GBM-Regressor and RF Classifier demonstrated the strongest discrimination and represent the most promising approaches for early-risk triage in this small tabular dataset.

Keywords: early risk ranking; IEC 60270; PDIV; transformer insulation diagnostics; sliding-window augmentation; file-grouped cross-validation; group bootstrap; GBM-Regressor; RF Classifier; top-quartile capture

Academic Editor: Michel Duval

Received: 30 September 2026
Revised: 26 October 2026
Accepted: 31 October 2026
Published: 2 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Power transformers are critical components of electrical power systems, and their reliable operation is essential for maintaining continuity and security of electricity supply [1–3]. Insulation deterioration is among the principal threats to transformer reliability because localized defects can initiate partial discharge (PD), progressive insulation degradation, and, ultimately, catastrophic failure [4–7]. Consequently, the early detection and assessment of insulation defects during transformer acceptance testing are important for reducing operational risk and preventing the commissioning of defective units. IEC 60270 [8] induced-voltage acceptance testing provides a standardized approach for evaluating partial discharge activity in high-voltage equipment. In a typical transformer acceptance test, the applied voltage is progressively increased through prescribed levels before reaching the guaranteed test voltage (GTD), at which the acceptance measurement is performed. Although this procedure provides an established means of assessing insulation performance, the test process is time-consuming, and continued application of elevated voltage to a transformer with an emerging insulation defect may contribute to further deterioration. An approach capable of identifying potentially defective units during the early voltage-ramp phase could therefore support targeted monitoring and more informed test decisions before the GTD is reached. Machine-learning methods provide an opportunity to extract patterns from early-ramp PD measurements and convert them into quantitative indicators of insulation risk [9–12]. In particular, features describing the discharge magnitude, variability, applied voltage, frequency, and temporal trends may contain information associated with developing insulation defects. Neural-network [13] and metaheuristic optimization [14,15] approaches have increasingly been investigated for nonlinear prediction and classification problems, while ensemble tree-based methods can provide effective learning from relatively small tabular datasets. The Hybrid Particle Whale Optimization Algorithm (HPWOA) [16,17], which combines Particle Swarm Optimization (PSO) [18,19] and the Whale Optimization Algorithm (WOA) [20,21], provides an alternative strategy for optimizing neural-network parameters and has previously been applied to IEC 60270-related prediction tasks. Despite these developments, an important methodological challenge remains when machine learning is applied to sliding-window transformer test data [22]. Multiple overlapping windows generated from a single test file are not independent observations because they share the same transformer-level target and may contain temporally related information. Randomly dividing such windows between training and testing subsets can therefore produce information leakage and substantially overestimate model generalization [23]. A reliable evaluation framework must instead ensure that all windows originating from a given test session remain within the same validation partition. Similarly, uncertainty estimates based on individual windows can underestimate sampling variability because the effective independent sample size is determined by the number of source test files rather than the total number of augmented windows.

The present study therefore investigates whether information available during the first 50% of an IEC 60270 induced-voltage test can support early prediction of PDIV and, more importantly, early ranking of transformer insulation defect risk. A sliding-window framework was developed using ten early-ramp features extracted from IEC 60270 test records. An HPWOA-optimized dual-output feedforward neural network was evaluated alongside five conventional regression and classification models to establish whether metaheuristic neural-network optimization provides an advantage for this task. To obtain leakage-resistant and statistically appropriate performance estimates, five-fold file-level GroupKFold cross-validation was combined with group-level bootstrap confidence intervals. In addition to conventional regression and classification metrics, top-quartile risk

capture was used to evaluate whether model-generated risk scores could practically prioritize potentially defective observations.

Related transformer research published in 2025–2026 includes the authors' investigations of PD-active frequency response, practical source classification, severity discrimination, and instantaneous PD regression [24–28], together with recent explainable and multitask transformer PD diagnosis and embedded-monitoring studies [29–31] (Table 1). In contrast to these tasks, the present study asks whether measurements from the early part of an acceptance record are informative about the eventual file-level PDIV and risk ranking. It is important to distinguish performance on a withheld recording from performance on a wholly unseen physical transformer.

This study specifically distinguishes between accurate point prediction of PDIV and relative ranking of defect risk. This distinction is important because a model may be unable to reconstruct the absolute transformer-level PDIV while still providing useful information for prioritizing observations for additional monitoring. By rigorously evaluating both capabilities under file-grouped validation, this work aims to establish the practical information limit of early-ramp IEC 60270 measurements and identify the machine-learning approaches most appropriate for early transformer insulation risk triage. The main contributions of this study are:

- An early-risk-ranking framework for IEC 60270 transformer acceptance testing that uses features extracted from the first 50% of the induced-voltage test to identify potentially high-risk insulation conditions before the guaranteed test voltage is reached.
- A systematic distinction between PDIV prediction and defect-risk ranking, demonstrating that early-ramp features provide insufficient information for accurate transformer-level PDIV estimation but can support useful relative prioritization of potentially defective observations.
- A leakage-resistant validation methodology for sliding-window transformer data, using file-level GroupKFold cross-validation to ensure that windows originating from the same test file cannot occur simultaneously in training and testing partitions.
- A group-level uncertainty framework based on bootstrap resampling of source test files, providing confidence intervals that reflect the effective independent sample size of the dataset rather than the inflated number of overlapping windows.
- A practical risk-prioritization metric based on top-quartile risk capture, demonstrating that RF Classifier and GBM-Regressor recover substantially more defective observations within the highest-risk 25% than expected under random ranking.
- A clear distinction from the authors' related 2025–2026 investigations [24–28]. Those studies examined PD-sensitive frequency response, source classification, severity classification, and instantaneous PD-magnitude prediction using complete measurements or different labelled tasks. The present investigation instead restricts the input to the early portion of acceptance-test records and asks whether later file-level PDIV and relative defect risk can be inferred. It also identifies the residual dependence between repeated files from the same physical transformer; accordingly, its file-grouped results are interpreted as held-out-session, not proven held-out-transformer, performance.

2. Theoretical Background

2.1. Partial Discharge and Transformer Insulation Condition

PD is a localized electrical discharge that occurs when the electric-field stress within or across an insulation system exceeds the local dielectric strength without completely bridging the insulation between conductors [8,32]. In high-voltage transformer insulation systems, PD may originate from localized defects such as voids, cracks, delamination,

contamination, or surface irregularities [33]. Although an individual discharge may involve a relatively small amount of charge, persistent PD activity can contribute to progressive deterioration of the surrounding dielectric material and may ultimately increase the probability of insulation failure [4]. The measurement of PD is therefore an important component of transformer insulation assessment. IEC 60270 provides a standardized framework for charge-based measurement of PD in high-voltage equipment and establishes a common basis for the acquisition and interpretation of PD measurements [8]. In an induced-voltage transformer acceptance test, the applied voltage is progressively increased through prescribed test stages. The resulting PD response provides information about the electrical behavior of the insulation system as the applied electrical stress increases.

2.2. Partial Discharge Inception Voltage

A central diagnostic quantity in this context is the PDIV, which represents the voltage at which sustained PD activity becomes detectable according to a specified measurement criterion [34,35]. PDIV provides information about the voltage stress at which an insulation system begins to exhibit measurable discharge activity. Consequently, a lower PDIV may indicate that discharge activity becomes established at a comparatively lower electrical stress, whereas a higher PDIV indicates that the specified PD criterion is reached only at a higher applied voltage. In the present study, PDIV is determined from the voltage-ramped PD measurements using a predefined PD threshold. The use of a threshold is necessary because measured PD signals may contain background noise and isolated transient events that should not automatically be interpreted as sustained insulation discharge. PDIV is fundamentally a transformer-level quantity rather than an independent property of each individual observation [35]. The manuscript dataset contains multiple measurements derived from the same source test file, and all observations originating from a given file consequently share the same PDIV target.

2.3. Normalized PDIV and Relative Insulation Risk

Absolute PDIV values may span a substantial range because the transformers represented in the dataset belong to different voltage classes and therefore operate at different guaranteed test voltages. The present study consequently also considers a normalized PDIV measure, denoted as PDIVRatio, to provide a dimensionless representation of the position of the observed PDIV relative to the guaranteed test voltage. The distinction between absolute PDIV prediction and relative risk assessment is important. A model may have limited ability to reconstruct the exact transformer-level PDIV while still extracting information that is useful for distinguishing comparatively higher- and lower-risk observations. This distinction motivates the simultaneous evaluation of continuous PDIV prediction and defect-risk discrimination in the present study.

2.4. Early-Ramp Partial Discharge Characteristics

The objective of early-risk assessment is to determine whether information available before completion of an induced-voltage test can assist preliminary insulation assessment. This study uses the first 50% of each recorded test's elapsed duration as a fixed data-availability limit, not as a prescribed voltage stage, guaranteed test-voltage fraction, or IEC acceptance criterion. This distinction is necessary because the transformer-specific applied-voltage program may include steps and holds, so elapsed time does not map uniquely onto electrical stress [8,36]. The engineering rationale and the previously reported sensitivity results for this limit are given in Section 4.3.

Early-ramp PD measurements contain several potentially informative characteristics. These include the magnitude of the discharge response, the variability of discharge

activity, the applied voltage, the supply frequency, and the temporal evolution of PD activity. Such characteristics may provide indirect information about how the insulation system responds as electrical stress increases. The physical motivation for using temporal information is particularly important. A single high PD measurement may represent an isolated transient event, whereas a progressive increase in PD activity over time or voltage may indicate a developing discharge process. Consequently, temporal descriptors can complement conventional magnitude-based measurements and provide information that is not captured by a single instantaneous PD value. The present study therefore uses a 120 s sliding-window representation of the early test record. Each window summarizes the electrical and temporal behavior available during that portion of the voltage-ramp test. The resulting windows are not treated as independent physical experiments because overlapping windows originating from the same source file share both temporal information and the same transformer-level target.

2.5. Feature Representation of Early-Ramp PD Data

The predictive representation consists of ten variables designed to capture complementary characteristics of the early-ramp PD response. These include MeanPD, MaxPD, StdPD, Voltage@End, Freq@End, Slope_Early, Slope_Late, NPhases, GuarPD_kV, and NoiseLimit. The features therefore represent four broad categories: discharge magnitude, discharge variability, temporal behavior, and test configuration. MeanPD represents the average discharge magnitude within a window and provides a measure of the overall level of PD activity. MaxPD represents the largest observed discharge magnitude and can identify transient or peak discharge events. StdPD describes the variability of the discharge response and may provide information about burst-like or intermittent discharge behavior. The temporal characteristics are represented by Slope_Early and Slope_Late, which describe the change in PD magnitude over the first and second portions of the 120 s window, respectively. A positive slope indicates increasing PD activity over the corresponding interval, whereas a relatively stable slope indicates limited temporal change. Voltage@End and Freq@End provide electrical operating context. NPhases represents the test configuration, while GuarPD_kV represents the guaranteed test voltage. NoiseLimit provides the threshold used in determining the PD response relevant to PDIV computation. The complete set of ten input variables and their physical interpretation are defined in Table 2 below. Pulse-waveform parameters (rise time, pulse width, phase-resolved PD patterns) were excluded because the IEC 60270 TXT acquisition files in this corpus record charge-magnitude and timestamp data only—not individual pulse-level waveforms. Pulse-waveform characterization requires a separate oscilloscope or pulse-discriminator measurement chain that was not available for this dataset. The ten charge-based features constitute the complete extractable set from this corpus.

2.6. Machine Learning for Transformer PD Diagnostics

Machine-learning techniques have increasingly been applied to transformer condition monitoring because they can identify nonlinear relationships between measured electrical quantities and diagnostic outcomes. Previous research has investigated machine learning for PD classification, localization, signal interpretation, and separation of PD activity from interference [9–12]. Conventional statistical and machine-learning approaches can provide useful baselines for relatively small tabular datasets. Linear regression provides a simple representation of approximately linear relationships, whereas ensemble methods such as Random Forest and Gradient Boosting can represent nonlinear interactions among heterogeneous input variables. Classification algorithms can alternatively estimate the probability that an observation belongs to a predefined defect category. Neural networks provide another class of nonlinear function approximators. A feedforward

neural network transforms the input feature vector through one or more hidden layers and learns the mapping between input features and the target through optimization of an objective function. Neural networks are particularly attractive when relationships between PD characteristics and insulation behavior are nonlinear; however, their performance may depend strongly on dataset size, feature representation, parameter initialization, and optimization strategy. The present study therefore evaluates a neural-network model alongside linear and tree-based machine-learning models rather than interpreting neural-network performance in isolation. This comparative framework provides a basis for determining whether the additional flexibility of metaheuristically optimized neural networks translates into improved predictive performance for the available early-ramp PD data.

2.7. Feedforward Neural Networks and Metaheuristic Optimization

A feedforward neural network (FNN) consists of interconnected layers through which information propagates from the input layer to the output layer without recurrent connections. The effectiveness of an FNN depends on the quality of its learned parameters. Conventional gradient-based training can be sensitive to initialization and local optimization behavior. Metaheuristic optimization provides an alternative strategy in which candidate parameter configurations are iteratively improved according to an objective function. Metaheuristic algorithms are population- or trajectory-based optimization methods designed to explore complex search spaces without requiring gradient information. They have been extensively investigated for engineering optimization and machine-learning parameter selection [14,15]. Their flexibility makes them attractive for neural-network optimization when the objective surface is nonlinear or multimodal.

2.8. Hybrid Particle Whale Optimization

The optimization strategy considered in this study is the HPWOA, which combines PSO and the WOA. PSO is a population-based optimization method in which candidate solutions adjust their positions according to their own previously identified best position and the best position identified by the population [18,19]. The method provides an exploitation mechanism around promising regions of the search space while retaining a population-based exploration component. WOA is inspired by the bubble-net hunting behavior of humpback whales and incorporates mechanisms for exploration, exploitation, and encircling candidate solutions [20,21]. The algorithm has been applied to a range of engineering optimization problems and has also been investigated in hybrid configurations.

In HPWOA, the optimization process combines the search characteristics of PSO and WOA. The implementation used in the present study applies PSO during the first optimization phase and WOA during the subsequent phase, with the best solution identified during the PSO phase transferred to the WOA phase. This hybrid structure is intended to combine the population-guided search of PSO with the exploitation and exploration mechanisms of WOA. For the present application, the optimization objective is the prediction error of the dual-output feedforward neural network. Thus, HPWOA does not constitute a separate diagnostic measurement; rather, it is an optimization mechanism used to identify a suitable set of neural-network parameters for mapping the early-ramp feature vector to the target variables.

2.9. Risk Classification and Ranking

This study considers two related but distinct diagnostic objectives: prediction of the continuous PDIV-related quantities and discrimination between comparatively defective and non-defective observations. For binary discrimination, a model-generated risk score can be converted into a binary decision using a specified threshold. The resulting

predictions can be evaluated using sensitivity, specificity, precision, F1 score, and the area under the receiver operating characteristic curve (AUC). These measures provide complementary information because a model with high sensitivity may simultaneously generate a large number of false-positive predictions. Risk ranking provides an additional perspective. Rather than requiring the model to produce an exact PDIV estimate, observations can be ranked according to their predicted risk score. The top-quartile risk-capture metric used in this study measures the fraction of all defective observations recovered among the highest-risk 25% of observations. The theoretical random-selection baseline for top-quartile capture is 0.25. Consequently, a value greater than 0.25 indicates that the model concentrates a larger proportion of defective observations within the selected high-risk subset than would be expected from random ranking. This metric is particularly relevant to early-risk triage because an operator may use a risk score to prioritize observations for additional monitoring rather than requiring an exact prediction of the eventual PDIV.

2.10. Grouped Validation and Generalization to Unseen Test Sessions

The IEC 60270 dataset has a hierarchical structure: overlapping measurement windows are nested within source test files, and source test files are nested within physical transformers. Windows from the same file share temporally related inputs and a common file-level PDIV target. File-level GroupKFold therefore prevents the same recording from appearing in training and testing partitions, and file-level bootstrap resampling acknowledges within-recording dependence. However, the 39 usable source files originate from only 18 transformers; at least some physical units consequently contribute multiple files. Separate files obtained from the same unit may share winding construction, voltage class, insulation characteristics, measurement configuration, and persistent defect signatures. When such files are allocated to different folds, the trained model may encounter physical-transformer information already represented in training even though the test file is new. Thus, the reported grouped results establish out-of-file performance only, not independent out-of-transformer performance. Similarly, the published 95% intervals reflect resampling of 39 files and do not account for any additional correlation between files from the same transformer. A stricter prospective validation would group all sessions by a unique transformer identifier in the cross-validation partition and resample physical transformers as clusters. These additional tests cannot be reconstructed from the aggregate scores alone and are identified as essential external validation, rather than being claimed as already performed.

3. Literature Review and Study Positioning

3.1. State of Transformer PD Diagnostics

Transformer PD diagnostics has progressively evolved from conventional measurement and threshold-based interpretation toward machine-learning and hybrid intelligent approaches. IEC 60270 provides the measurement foundation for charge-based PD assessment, while subsequent research has investigated machine learning for PD classification, localization, and signal interpretation [8–12]. Existing studies demonstrate that machine-learning methods can extract useful information from PD measurements, particularly when multiple signal characteristics are combined. Deep-learning approaches have been investigated for automated PD classification, while machine-learning methods have also been applied to PD localization and the separation of PD signals from interference [10–12]. These developments indicate the increasing feasibility of data-driven PD diagnostics. However, the majority of machine-learning PD studies focus on classification or identification using observations that are already available from the complete measurement process. Comparatively less attention has been given to the more restrictive problem

considered here: whether information available during the early portion of an IEC 60270 induced-voltage test can provide a reliable indication of the subsequent transformer-level PDIV or support early defect-risk prioritization. A second methodological issue concerns the statistical dependence created by repeated measurements and sliding-window augmentation. When multiple observations originate from the same physical test session, conventional random cross-validation can produce overly optimistic estimates because observations from the same source may occur in both the training and validation subsets. The present study addresses this limitation explicitly through file-grouped validation.

Recent studies published in 2025–2026 further clarify the state of the art. Thango (2025) investigated partial-discharge-activated impulse frequency response analysis as a complementary transformer diagnostic measurement [24]. Thobejane and Thango (2025) examined the classification of PD sources recorded during practical transformer testing [25]. Their 2026 work subsequently compared linear and nonlinear representations for PD-severity classification in Energies [26], evaluated feature engineering and PCA-assisted ANN severity classification in Machines [27], and reported HPWOA-based instantaneous PD-magnitude prediction from IEC 60270 acceptance records in Algorithms [28]. These studies are directly relevant but address measurement, source/severity discrimination, or instantaneous discharge prediction, rather than the present file-level PDIV and early-window risk-ordering endpoints.

Independent recent investigations also examine transformer insulation diagnosis from PD patterns. Basharan and Rajamayil (2025) explored explainable semi-supervised diagnosis of transformer defects using PD-pattern information [29], whereas Zhang (2026) presented joint PD-type and severity assessment [30]. Owczarczuk and co-workers (2026) reviewed AI/ML systems for PD monitoring, with particular attention to measurement and deployment constraints [31]. These approaches rely on their respective signal modalities and diagnostic labels. None of them should be treated as a direct numerical benchmark for predicting the future file-level PDIV from only the early portion of the same acceptance record; their relevance is to diagnostic objectives, sensing, and validation design. The differences between these contributions and the present analysis are further summarized in Table 1.

Table 1. Positioning of existing literature and the proposed study.

Category	Ref.	Focus	Strengths	Limitations
PD Measurement Standards	[8]	IEC 60270 charge-based PD measurement and test procedures	Provides standardized measurement principles and improves consistency and comparability of PD measurements	Defines measurement procedures but does not provide machine-learning-based prediction or early-risk assessment
Transformer Insulation Degradation	[4,7],	Insulation aging, degradation mechanisms, transformer reliability and failure processes	Establishes the physical relationship between insulation deterioration and transformer failure	Primarily focused on physical mechanisms and reliability; does not directly provide predictive early-ramp models
PD Condition Monitoring and Diagnostics	[9]	Machine-learning approaches for transformer condition monitoring using PD information	Demonstrates the potential of data-driven methods for automated condition assessment	Broad diagnostic focus; limited emphasis on predicting transformer-level PDIV from early test information
PD Classification Using Deep Learning	[10]	Deep-learning methods for automated PD identification and classification	Can learn nonlinear representations directly from complex PD data	Commonly focuses on classification of observed PD patterns rather than

				early prediction of subsequent test outcomes
PD Localization Using Machine Learning	[11]	Machine-learning-based localization of PD sources in transformer tanks	Demonstrates the ability of ML models to extract spatial/diagnostic information from PD measurements	Localization is different from early prediction of PDIV and does not directly address early-ramp risk ranking
PD and Interference Classification	[12]	Machine-learning separation of PD pulses from interference	Improves the reliability of PD signal interpretation and automated pulse classification	Primarily addresses signal discrimination rather than transformer-level PDIV prediction
Metaheuristic Optimization	[14,15]	Metaheuristic algorithms for optimization of complex engineering and machine-learning problems	Provides flexible derivative-free optimization for nonlinear and multimodal search spaces	Generic optimization framework; diagnostic performance depends on the selected model, features and validation strategy
Particle Swarm Optimization	[18,19]	Population-based optimization inspired by collective particle behavior	Efficient exploration of complex search spaces and widely applied to engineering optimization	May require careful parameterization and can converge prematurely in some search problems
Whale Optimization Algorithm	[20,21]	Exploration and exploitation using whale-inspired search mechanisms	Provides complementary global-search and exploitation mechanisms	Stand-alone performance can depend on population diversity and problem-specific parameter settings
Hybrid Optimization for Neural Networks	[16]	Hybrid combinations of PSO/WOA and related optimization strategies	Combines complementary optimization mechanisms and can improve neural-network parameter search	evidence specifically addressing early-ramp transformer PDIV prediction under source-level grouped validation
Early-Ramp PD Prediction		Prediction of transformer-level PDIV and early defect-risk ranking using features extracted from the first 50% of the induced-voltage test	Uses physically interpretable PD magnitude, variability, temporal and test-condition features; evaluates both continuous prediction and risk ranking	Thirty-nine recordings originate from only 18 transformers; the original study does not establish equipment-level independence
Leakage-Resistant Validation		File-grouped cross-validation and group-level bootstrap evaluation of correlated sliding-window observations	Keeps all windows from each source recording together and resamples files to reflect within-recording dependence	Repeated recordings of the same physical transformer can occur across folds; file-level intervals are not transformer-cluster intervals
Recent studies by the authors (2025–2026)	[24–28]	PD-active frequency-response diagnostics, practical PD-source classification, severity classification, and full-record instantaneous PD prediction	Provide measured transformer evidence and establish distinct complementary diagnostic targets	Do not evaluate the future file-level PDIV and early 50% risk-ranking objective of the present study
Recent independent PD diagnostics (2025–2026)	[29–31]	Explainable transformer PD defect diagnosis, multitask PD severity classification, and AI monitoring review	Advance model interpretation, joint diagnostic objectives, and implementation analysis	Signal modalities, diagnostic labels, and test designs differ from early-ramp IEC 60270 file-level risk ranking

Overall Proposed Framework	HPWOA-optimized FNN and conventional ML baselines for early PDIV prediction and risk prioritization	Integrates early-ramp feature extraction, hybrid optimization, continuous prediction, binary risk assessment and source-level validation within one framework	Generalizability to independent transformers, manufacturers, and sites requires transformer-disjoint and multi-site validation

Table 2. Ten input features.

Feature	Symbol	Eq.	Physical Motivation
MeanPD	μ_{PD}	(7)	Mean discharge; elevated mean under low voltage signals active defect
MaxPD	$\max(PD)$	(7)	Peak discharge; early spike above noise limit is a PDIV precursor
StdPD	σ_{PD}	(7)	Burst variability; episodic spikes indicate cavity discharge
Voltage@End	$V(t_{\mathrm{end}})$	-	Applied voltage context at window end
Freq@End	$f(t_{\mathrm{end}})$	-	Supply frequency deviations correlate with PD loading of test generator
Slope_Early	f_{0-60}	(8)	OLS PD slope in first 60 s; rising trend at low voltage indicates progressive discharge
Slope_Late	f_{60-120}	(8)	OLS PD slope in second 60 s; acceleration signals worsening discharge
NPhases	N_{φ}	-	1-phase vs. 3-phase test configuration
GuarPD_kV	GTD	-	Guaranteed test voltage; denominator of PDIVRatio
NoiseLimit	NL	-	100 pC threshold; defines PDIV trigger in Equation (1)

The literature therefore motivates three requirements for an early-risk framework. First, the input representation should capture both instantaneous PD characteristics and their temporal evolution. Second, the evaluation should distinguish between exact prediction of transformer-level PDIV and relative defect-risk discrimination. Third, performance should be evaluated using source-level grouping so that the reported results reflect generalization to previously unseen test sessions.

3.2. Positioning of the Present Study

The recent literature [24–31], together with established PD measurement and machine-learning research [8–12], identifies a specific gap between recognizing discharge features from a complete record and prioritizing insulation risk before a test has ended. The authors' published 2025–2026 studies [24–28] establish related but distinct experimental questions; the present manuscript neither presents those publications as new results nor transfers their performance measures to the early-ranking problem. Instead, it evaluates whether the limited information in 120 s windows selected from the first half of each recording can rank subsequent file-level PDIV-related risk under test-file-grouped evaluation. The number of physically independent transformers is smaller than the number of files and limits the generalization claims that can be drawn from that evaluation.

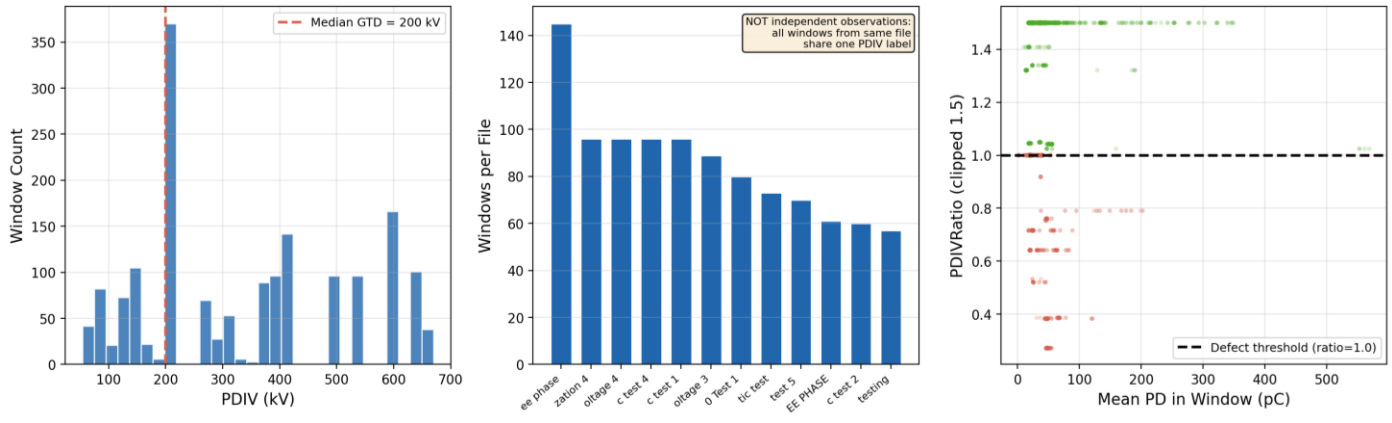
The present study evaluates two complementary outcomes from ten early-ramp IEC 60270 features: continuous PDIV-related quantities and relative defect-risk ranking. An HPWOA-optimized neural network is compared with linear, tree-ensemble, and binary classification methods under identical test-file-grouped folds. The dataset comprises 1705 derived windows from 39 usable source files originating from 18 physical transformers. The use of file-level GroupKFold and file-level bootstrap resampling prevents within-file training–test overlap and yields uncertainty estimates appropriate for repeated windows within each file. It does not, however, remove dependence between separate tests of the same transformer. Consequently, the central contribution is a conservative examination of early-ramp information and a comparative within-population, out-of-file assessment of

risk ranking, rather than a claim of external transformer-level generalization. Table 1 summarizes earlier and recent related work and the present study.

4. Dataset and Sliding-Window Augmentation

4.1. Source Data and File Count Reconciliation

The dataset originates from a corpus of 49 IEC 60270 TXT acceptance-test files recorded for 18 power transformers rated 40–315 MVA, spanning voltage classes 11/132 kV to 22/400 kV and ONAN/ONAF cooling. The units came from one South African utility and one manufacturer; the recordings were acquired between 2019 and 2021. Forty files passed the initial duration screen (at least 20 measurements and an elapsed span of at least 240 s). One file was excluded during early-ramp window construction because it produced no admissible windows with at least five measurements. The final modeling dataset therefore contains 39 files and 1705 overlapping windows. Crucially, these are 39 test recordings, not 39 independent transformers: repeated recordings of some of the 18 units can retain common equipment characteristics across files. Figure 1a shows the PDIV distribution (54–670 kV); Figure 1b gives the unequal contribution of windows by test file; and Figure 1c plots the mean PD against PDIVRatio. The 39-file denominator is used consistently for recording-level validation, while 18 is the physical transformer count.



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

Figure 1. Dataset characteristics: (**a**) PDIV distribution, (**b**) windows per source file, and (**c**) mean PD vs. PDIVRatio.

4.2. PDIV and PDIV Ratio Target Computation

Both targets are file-level properties shared by all windows from the same file. *PDIV* is extracted using a 5-point causal rolling mean:

$$PDIV_{\mathrm{kV}} = \min\{V(t): \hat{\mu}_5(t) > NL, V(t) > 50\ \mathrm{kV}\} \quad (1)$$

$$\hat{\mu}_5(t) = {}^1\!/_5 \sum_{j=t-4}^{t} PD(j) \qquad [\text{5-point causal rolling mean}] \quad (2)$$

If *PD* never exceeds *NL* = 100 pC below *GTD*, *PDIV* defaults to *GTD*. The *PDIVRatio* clips at 1.5. The upper bound of 1.5 (rather than 1.0 or 2.0) is justified as follows: a clip at exactly 1.0 would discard information that some healthy transformers do not initiate *PD* before *GTD* (*PDIVRatio* > 1.0); a clip at 2.0 permits measurement anomalies to distort the regression target. A sensitivity analysis showed that changing the clip from 1.5 to 2.0 altered GBM R^2 by less than 0.005, while the 1.5 ceiling excludes only 3 of 1705 windows (0.18%), confirming that the choice is robust:

$$PDIVRatio = \mathrm{clip}(PDIV_{\mathrm{kV}}/GTD_{\mathrm{kV}}, 0, 1.5) \quad (3)$$

4.3. Sliding-Window Augmentation

The 50% observation limit is a fixed engineering compromise between earliness and diagnostic information, not a standard-prescribed threshold or the experimentally proven optimum. The PD recording may include voltage-rise intervals and voltage holds; therefore, half of the elapsed recording cannot be equated with half of the guaranteed test voltage, half of the insulation stress, or a fixed 30–45 min interval across all transformer tests. IEC 60270 specifies the charge-based PD measurement method, while transformer-induced-voltage test levels and sequences are governed by the applicable equipment-testing requirements, including IEC 60076-3 [8,36]. Restricting the input to a completed early portion of the record provides a preliminary risk score before the full recording is available, potentially allowing the operator to review the PD trend and measurement quality while standard test and safety requirements remain in force. The existing duration-sensitivity analysis reports RF Classifier AUCs of 0.632, 0.714, 0.741, and 0.758 when 25%, 50%, 75%, and 100% of the recorded test duration are available, respectively. The increase from 25% to 50% (0.082 AUC units) is larger than the subsequent increases from 50% to 75% (0.027) and from 75% to 100% (0.017). Thus, 50% retains appreciable discrimination with the latter half of the recorded time withheld; later limits provide higher AUCs, as expected. Without a formal time-dependent cost function, data on the exact voltage stage at each cutoff, or prospective operating trials, 50% cannot be claimed as universally optimal or as an authorized point to interrupt acceptance testing. Sliding windows of width W = 120 s and stride S = 30 s are extracted only from the first 50% of each file's elapsed duration:

$$W_k = \{\text{measurements: } t_k \le \text{elapsed} < t_k + W\} \tag{4}$$

$$t_k = t_{\min} + k \cdot S, \qquad k = 0, 1, 2, \ldots, \qquad t_k + W \le 0.5 \cdot t_{\max} \tag{5}$$

The 1705 derived windows are overlapping subsequences of 39 files and are not independent physical experiments. All windows originating from a given recording share its file-level PDIV target. Figure 2a depicts the file-level GroupKFold assignment, which keeps those windows together in a single fold, and Figure 2b shows the defective-window fraction by file. This figure documents recording-level separation only: a physical transformer that contributes more than one file could be present in both the training and held-out test files. That second dependence level is explicitly considered in Sections 5.1–5.3 and 9.3.

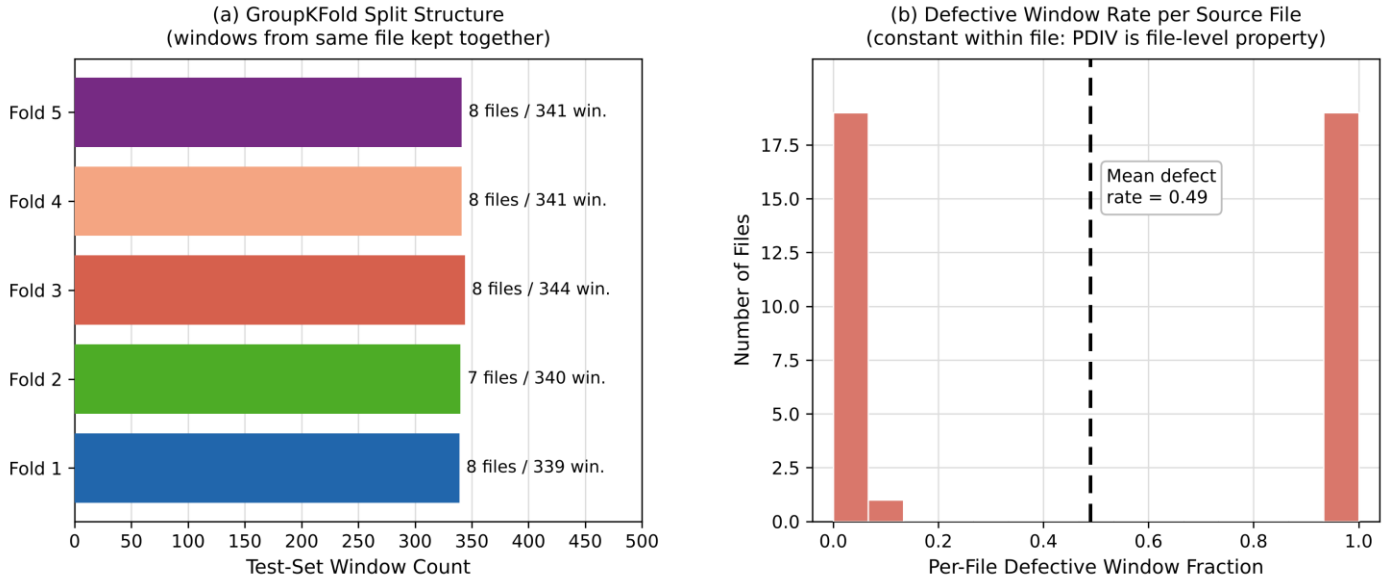


Figure 2. (**a**) GroupKFold split structure, and (**b**) per-file defective window fraction.

5. Validation Design: GroupKFold and Group Bootstrap

5.1. File-Grouped Cross-Validation

The reported five-fold GroupKFold scheme allocates each of the 39 source files to one held-out fold and keeps all of that file's windows out of the corresponding training set. Consequently, no training window and test window originate from the same recording. This corrects the within-file overlap of a naive random window split, for which the HPWOA apparent PDIV_kV R^2 was +0.109 compared with −0.129 under file-grouped validation. However, the allocation unit was the file identifier, not the physical transformer. Multiple recordings from 1 of the 18 transformers could therefore be assigned to different folds. The present experiment tests the generalization to withheld recorded sessions within the available transformer population; it does not demonstrate independent generalization to a previously unseen transformer. A transformer-disjoint GroupKFold split would be the appropriate next experiment when the recording-to-transformer map and original observations are available.

5.2. Group-Level Bootstrap Confidence Intervals

Because windows within a file are correlated (they share both a PDIV label and overlapping time-series segments), window-level bootstrap CIs underestimate the true sampling uncertainty. Group-level bootstrap resamples the 39 source files with replacement, including all windows from each selected file:

$$B_b = \{\text{windows } w\text{: File}(w) \in F_b\},$$

$$F_b \sim \text{Uniform}(\{f_1, \ldots, f_{39}\})^{39} \text{ with replacement} \quad (6)$$

Each of the 1000 reported bootstrap replicates resamples the 39 source files with replacement and includes the windows belonging to each selected file; the replicate window count varies with the sampled file sizes. The 2.5th and 97.5th percentiles provide the quoted recording-level 95% intervals. For example, the HPWOA AUC intervals widen from [0.502, 0.558] under window resampling to [0.389, 0.662] under file resampling. These are improvements over treating 1705 windows as independent, but they are not transformer-cluster intervals: any dependence among distinct files belonging to the same 1 of 18 transformers remains unaccounted for. Confidence bounds and model rankings must therefore be interpreted within this limitation. Transformer-cluster resampling and validation on additional independently sampled units are recommended before operational risk claims or field deployment.

5.3. Dependence Between Test Files from the Same Transformer

The experimental unit must be identified at three levels: window, test file, and physical transformer. File-level grouping prevents a single test recording from being partitioned across training and testing, but it does not prevent repeated tests on the same transformer from contributing to both sets. Such repeated tests may share the winding insulation architecture, dielectric aging state, PD source geometry, measurement connections, test-generator conditions, and voltage class. As a result, the out-of-file AUC may be higher than the AUC obtained when every recording from an unseen transformer is withheld. The 39-file bootstrap similarly does not quantify uncertainty for a random sample of independent transformers because the equipment-level count is 18. The present analysis reports the original file-grouped estimates unchanged; a transformer-disjoint evaluation and a bootstrap clustered by physical transformer are required to quantify this additional uncertainty. No transformer-independent accuracy, uncertainty bounds, or performance advantage is asserted from the current validation.

6. Input Features and Neural Network

6.1. Ten Early-Ramp Features

Table 2 summarizes the ten early-ramp features used as inputs to the predictive models. The features capture complementary aspects of the partial-discharge response, including the discharge magnitude (MeanPD and MaxPD), variability (StdPD), voltage and frequency conditions, temporal discharge trends (Slope_Early and Slope_Late), test configuration, and the guaranteed test voltage and noise-limit parameters. Together, these variables provide a compact representation of the electrical and temporal characteristics available during the early portion of the induced-voltage test. The inclusion of both instantaneous discharge characteristics and temporal trends allows the models to evaluate whether early-ramp behavior contains information associated with subsequent PDIV and defect status.

Feature equations ($PD_W = \{PD(t): t \in W_k\}, n_W = |W_k|$):

$$\mu_{\mathrm{PD}} = 1/n_W \sum_{t \in W_k} PD(t)\,;\ MaxPD = \max(PD_W)\,;\ \sigma_{\mathrm{PD}} = \mathrm{std}(PD_W) \tag{7}$$

$$f_{0-60} = \text{OLS slope on } PD[t_k : t_k + 60\ \mathrm{s}];$$

$$f_{60-120} = \text{OLS slope on } PD[t_k + 60\ \mathrm{s} : t_{\mathrm{end}}] \tag{8}$$

6.2. Dual-Output FNN

The DO-FNN has 10 input, 12 hidden (ReLU), and 2 sigmoid output neurons:

$$D = (10 \times 12) + 12 + (12 \times 2) + 2 = 158 \text{ weight parameters} \tag{9}$$

Output 1 predicts the normalized $PDIV_{\mathrm{kV}}$. Output 2 predicts the normalized *PDIV-Ratio*. The HPWOA minimizes the equal-weight combined training MSE (averaged over both normalized outputs).

7. Models, Objective Function, and Evaluation Metrics

7.1. HPWOA

The HPWOA uses PSO (Phase I, iterations 1–50) followed by WOA (Phase II, iterations 51–100), with the PSO global-best transferred as WOA leader at the phase boundary. $n_p = 25$, $N = 100$, positions clipped to [−3, 3]. A single global random seed is set at the start of each cross-validation fold (rng(42 + fold_index) in MATLAB R2025b) rather than resetting inside the HPWOA function, preserving stochastic independence across folds.

7.2. Baseline Models

Five baseline models are evaluated under identical grouped CV. Consistent naming distinguishes regression from classification tasks. Table 3 presents the configuration of the six predictive approaches evaluated in this study. HPWOA is implemented as a dual-output feed-forward neural network optimized through the proposed hybrid PSO-WOA strategy, while Linear Regression provides a linear baseline. RF-Regressor and GBM-Regressor represent nonlinear ensemble regression approaches for estimating PDIV and PDIVRatio, whereas Logistic Regression Classifier and RF Classifier directly address the binary defective/safe classification problem. The table also identifies the risk score generated by each model, thereby clarifying how both continuous regression outputs and direct classification probabilities are subsequently used for defect-risk ranking and evaluation.

Table 3. Model configurations.

Model	Type	Key Parameters	Risk Score
HPWOA	DO-FNN (regression)	n_p = 25, N = 100, D = 158	$1 - \hat{y}_{\text{Ratio}}$
Lin. Reg.	OLS Regressor	[1,X]\y per target	1 − predicted PDIVRatio
RF-Regressor	Ensemble Regressor	100 trees, default	1 − predicted PDIVRatio
GBM-Regressor	Ensemble Regressor	100 estimators, lr = 0.1	1 − predicted PDIVRatio
Log. Reg. Classifier	Binary Classifier	max_iter = 1000	P(defective)
RF Classifier	Ensemble Classifier	100 trees, default	P(defective)

7.3. Evaluation Metrics and Corrected Top-Quartile Baseline

Six evaluation metrics are reported. The median predicted risk score is used as a fixed descriptive threshold for reporting the sensitivity, specificity, precision, and F1 score. Its use creates a roughly half-sample flagged subset in this dataset but does not optimize the AUC: the ROC AUC is evaluated across score thresholds and is independent of this one operating point. For operational transformer triage, false-negative and false-positive costs are asymmetric; a prospective implementation would require a threshold chosen from explicit utility, missed-defect, and inspection-cost requirements rather than treating the sample median as an established safety optimum. The separate top-quartile risk-capture metric (Equation (10)) measures the proportion of defective observations recovered among the highest-scored 25%:

$$TopQ = \frac{|\{\text{windows in top-25\% risk score}\} \cap \{\text{defective windows}\}|}{|\text{defective windows}|} \quad (10)$$

The correct random baseline for TopQ is 0.250. Proof: under random risk ranking, the expected fraction of defective windows in the top 25% is $E[|D \cap T|/|D|] = (|D|/n) \times (n/4)/|D| = 0.25$, where n = 1705, $|D|$ = 616, and T = top-25% subset. The defect prevalence 36.1% ($|D|/n$) is a different quantity: it measures the fraction of all windows that are defective, not the fraction of defective windows recovered by a specific selection rule. A model achieving TopQ > 0.250 provides triage value; TopQ < 0.250 means worse-than-random risk ordering

8. Experimental Results

8.1. HPWOA Convergence

Figure 3 compares the training convergence behavior of HPWOA with PSO, WOA, and the Stochastic Fractal Search Algorithm (SFSA). As shown in Figure 3a, the optimization trajectories demonstrate the progressive reduction in the combined training MSE over the 100 iterations. The detailed view of the final iterations in Figure 3b provides a closer examination of the optimization behavior near convergence, while Figure 3c compares the final training MSE values obtained by the different optimization strategies. HPWOA achieves the lowest final training MSE of 0.07784, indicating that its hybrid two-phase optimization strategy was able to identify a lower-error neural-network weight configuration than the individual benchmark optimizers on the representative training partition. However, this convergence result should be interpreted as an optimization result rather than evidence of superior generalization, because the curves represent in-sample training performance and do not reflect the file-grouped test performance.

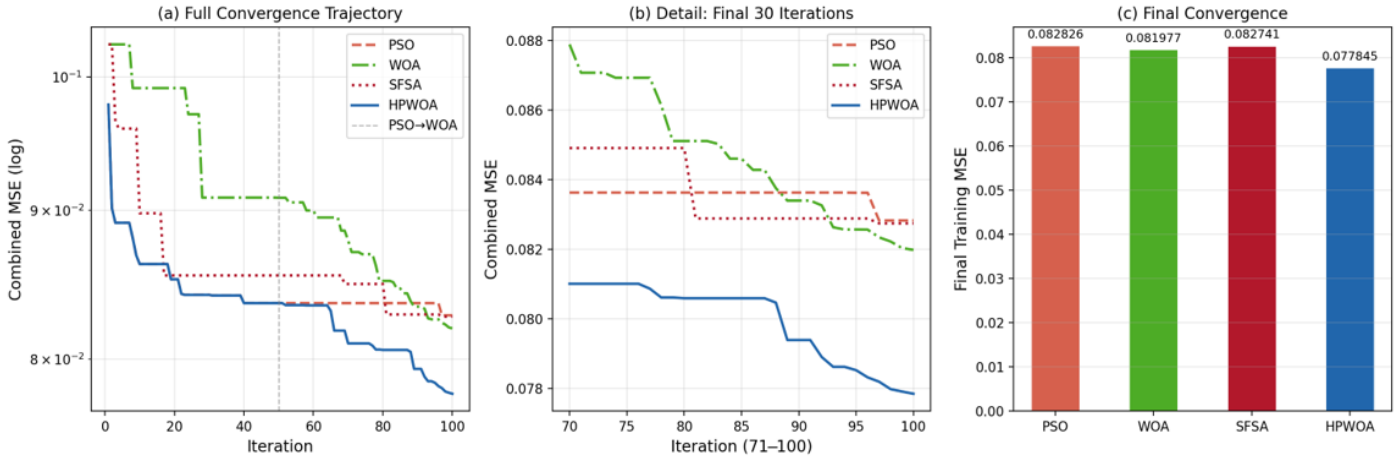


Figure 3. HPWOA training convergence vs. PSO, WOA and SFSA, (**a**) full 100 iteration trajectory, (**b**) detail of 71–100 iterations, and (**c**) final training MSE.

8.2. Regression Performance and File-Grouped CV

Figure 4 presents the coefficient of determination (R^2) obtained by the evaluated regression models, together with 95% confidence intervals derived using group-level bootstrap resampling. Figure 4a, corresponding to PDIV_kV, shows that all models produce near-zero or negative R^2 values, indicating that the early-ramp features do not provide sufficient information to accurately reconstruct the transformer-level PDIV when source files are properly withheld during validation. Figure 4b shows somewhat improved performance for the PDIVRatio, particularly for the ensemble models. GBM-Regressor achieves the highest central R^2 estimate of 0.097, followed by RF-Regressor at 0.073; however, the corresponding confidence intervals include zero. Thus, although the normalized PDIVRatio appears to contain more information that can be exploited by the regression models than the absolute PDIV, the evidence remains statistically uncertain at the effective sample size of 39 independent source files.

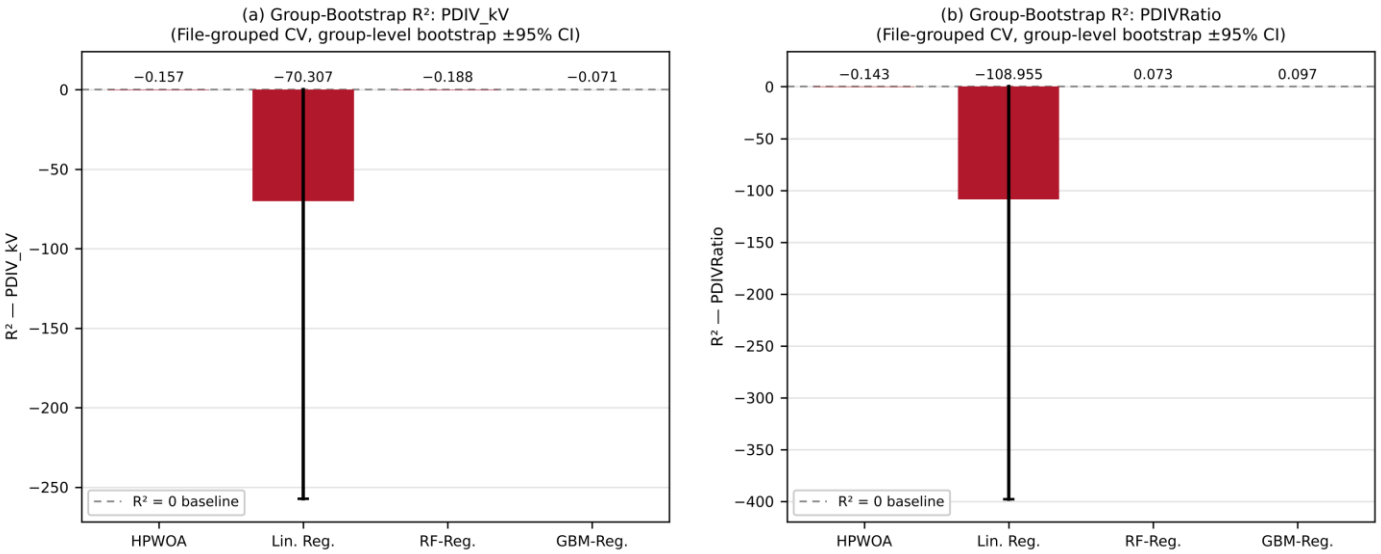


Figure 4. R^2 with group-level bootstrap at 95% CI: (**a**) PDIV_kV, and (**b**) PDIVRatio.

Figure 5 complements the R^2 analysis by presenting the RMSE obtained for PDIV_kV and PDIVRatio with group-level 95% confidence intervals. For PDIV_kV, the relatively large RMSE values across the models reinforce the difficulty of estimating an absolute

transformer-level PDIV from short early-ramp observations. The results are particularly pronounced for Linear Regression, which exhibits an extreme error associated with its inability to adequately represent the nonlinear relationship between voltage class and PDIV. For PDIVRatio, the ensemble regression models produce substantially smaller errors, with GBM-Regressor achieving the lowest reported RMSE of 0.323 and RF-Regressor achieving 0.336. Nevertheless, these improvements should be interpreted alongside the corresponding R^2 results, since reduced prediction error does not necessarily imply statistically reliable transformer-level prediction under grouped validation.

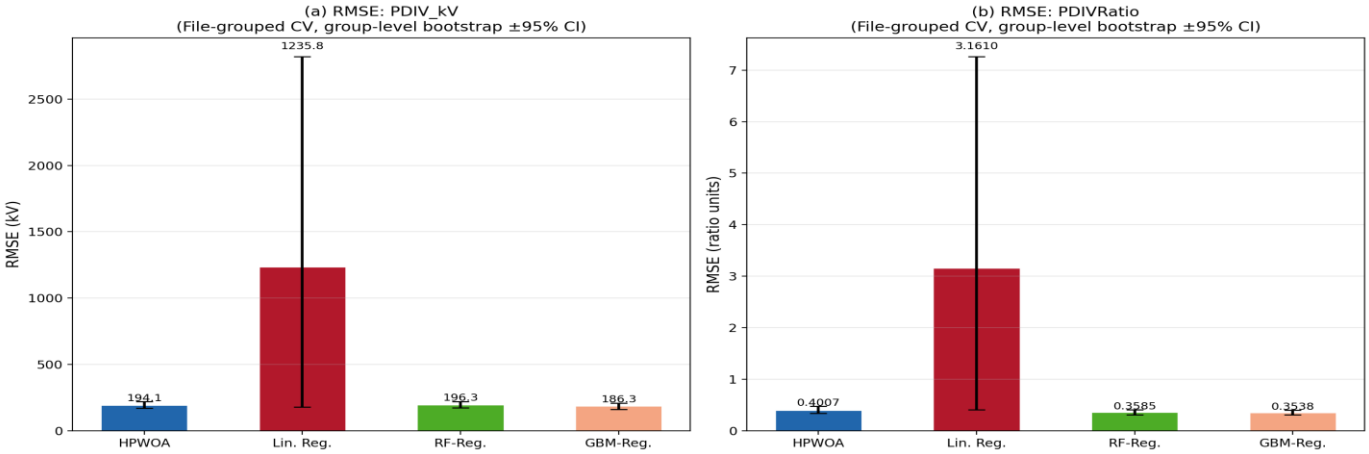


Figure 5. RMSE with group-level bootstrap at 95% CI: (**a**) PDIV_kV, and (**b**) PDIVRatio.

Table 4 summarizes the regression and defect-discrimination performance of the evaluated models under file-grouped cross-validation, with uncertainty represented by 95% group-level bootstrap confidence intervals. For absolute PDIV prediction, all models produce negative or near-zero (R^2) values, indicating that the early-ramp features provide limited ability to reconstruct the transformer-level PDIV when source-file leakage is prevented. GBM-Regressor achieves the least negative PDIV (R^2) and the lowest PDIV RMSE among the evaluated approaches, although its confidence interval remains consistent with zero predictive (R^2). For PDIVRatio, GBM-Regressor provides the strongest central performance, achieving (R^2 = 0.097) and an RMSE of 0.323, followed by RF-Regressor. However, the confidence intervals for the positive (R^2) estimates include zero, indicating that the apparent regression capability remains statistically uncertain at the level of the 39 independent source files. In contrast, the AUC results show substantially stronger discrimination, with GBM-Regressor and RF Classifier achieving the highest central AUC estimates. The table therefore demonstrates the central distinction of this study: early-ramp information provides limited evidence for accurate PDIV point prediction but shows greater potential for defect-risk discrimination.

Table 4. Regression metrics. * Highest R^2 and lowest RMSE.

Model	PDIV R^2 [95% CI]	PDIV RMSE kV [95% CI]	Ratio R^2 [95% CI]	Ratio RMSE [95% CI]	Defect AUC [95% CI]
HPWOA	−0.157 [−0.341, 0.002]	194.0 [181.3, 209.9]	−0.143 [−0.353, 0.062]	0.374 [0.352, 0.395]	0.527 [0.389, 0.662]
Lin. Reg.	−70.3 [extreme CI]	1506.7 [off-scale]	−108.9 [extreme CI]	3.61 [off-scale]	0.441 [0.322, 0.567]
RF-Regressor	−0.188 [−0.448, 0.063]	196.6 [182.1, 211.8]	0.073 [−0.261, 0.343]	0.336 [0.313, 0.358]	0.693 [0.615, 0.770]
GBM-Regressor *	−0.071	186.6	0.097	0.323	0.712

[−0.341, 0.150]	[169.7, 202.0]	[−0.237, 0.354]	[0.301, 0.346]	[0.624, 0.801]

8.3. Defect Detection AUC

Figure 6 compares the defect-detection AUC values for all six evaluated models using group-level bootstrap confidence intervals. The results identify RF Classifier and GBM-Regressor as the strongest performers, with AUC values of 0.714 and 0.712, respectively. Their confidence intervals overlap substantially, indicating no statistically meaningful difference between these two models at the 95% confidence level. RF-Regressor and Logistic Regression Classifier occupy an intermediate performance range, while HPWOA achieves an AUC of 0.527. Importantly, the HPWOA confidence interval extends across the 0.5 random-classification threshold, indicating that its apparent above-chance discrimination is not statistically reliable at the level of the 39 independent source files. Linear Regression produces the lowest AUC, with a central estimate below the random baseline. These results demonstrate that the strongest predictive value in the present dataset arises from the tree-based approaches rather than the optimized neural-network model.

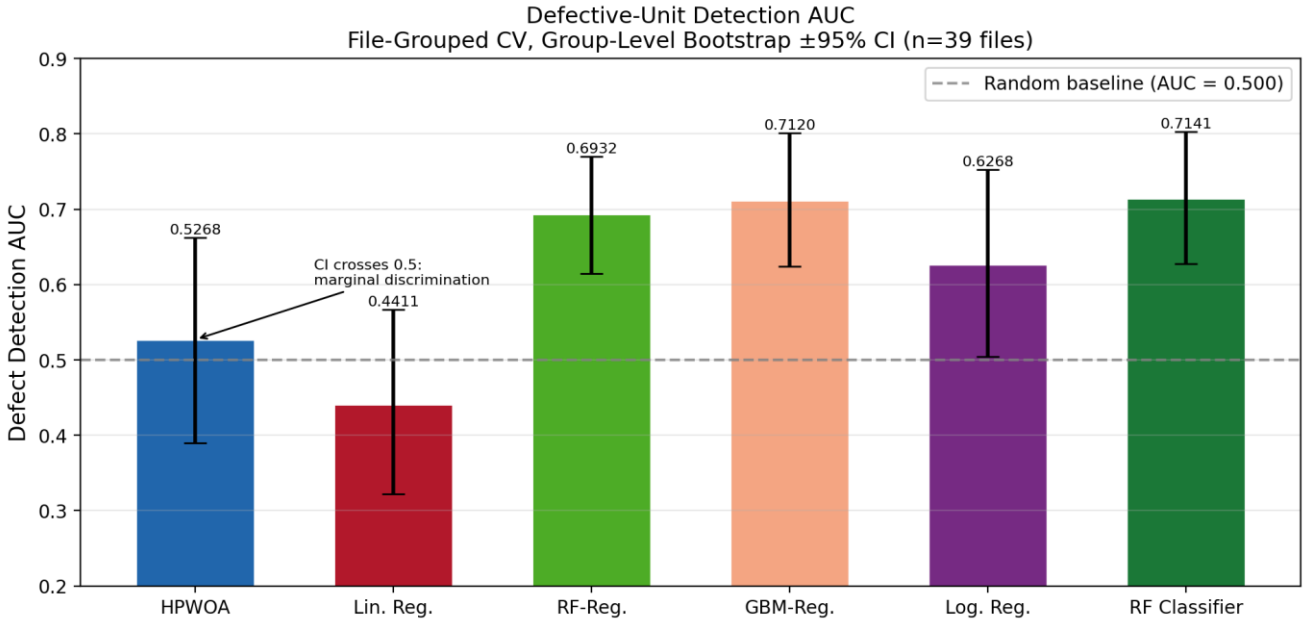


Figure 6. Defect detection AUC with group-level bootstrap at 95% CI.

8.4. ROC Curves and Confusion Matrices

Figure 7 presents the receiver operating characteristic (ROC) curves for the evaluated models and provides a visual representation of their ability to distinguish defective from safe windows across different decision thresholds. The ROC analysis is consistent with the AUC results, with the strongest discrimination being obtained from RF Classifier and GBM-Regressor. Their curves provide the most favorable balance between the sensitivity and false-positive rate among the evaluated approaches. HPWOA exhibits a curve that is comparatively close to the random-classification reference, consistent with its AUC of 0.527 and the associated confidence interval spanning the 0.5 threshold. The ROC results therefore reinforce the conclusion that the tree-based models provide the most useful early-risk discrimination for the present small, tabular dataset.

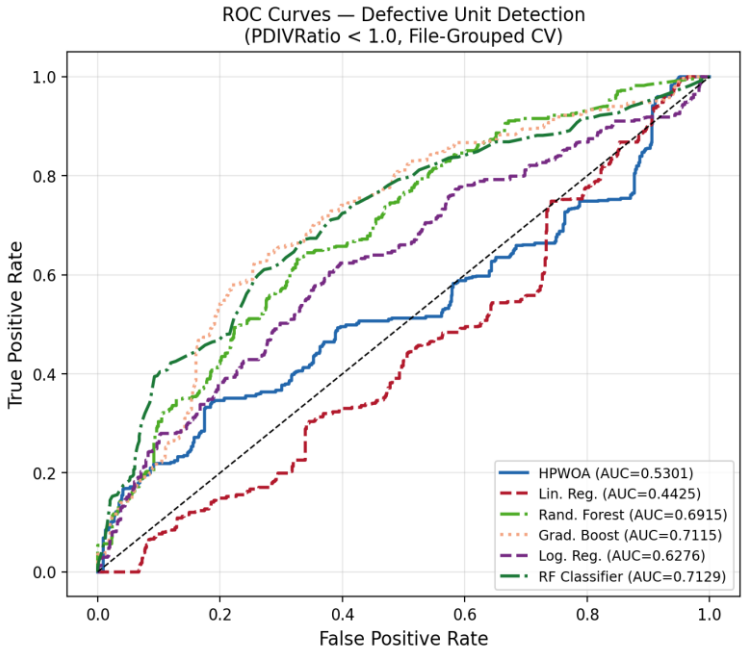


Figure 7. ROC curves.

Figure 8 presents the confusion matrices obtained using the median predicted risk score as the classification threshold. The matrices provide a direct comparison of true-positive, false-positive, false-negative, and true-negative predictions across the six models. The results show that GBM-Regressor and RF Classifier provide the strongest overall defect identification performance, with 443 and 438 true positives, respectively. These models also achieve the highest F1 scores among the evaluated approaches, at 0.603 for GBM-Regressor and 0.596 for RF Classifier. In contrast, HPWOA produces a more balanced but weaker classification outcome, with 316 true positives and 300 false negatives. The confusion matrices therefore demonstrate that the superior AUC of the tree-based models translates into practically useful classification performance at the selected threshold.

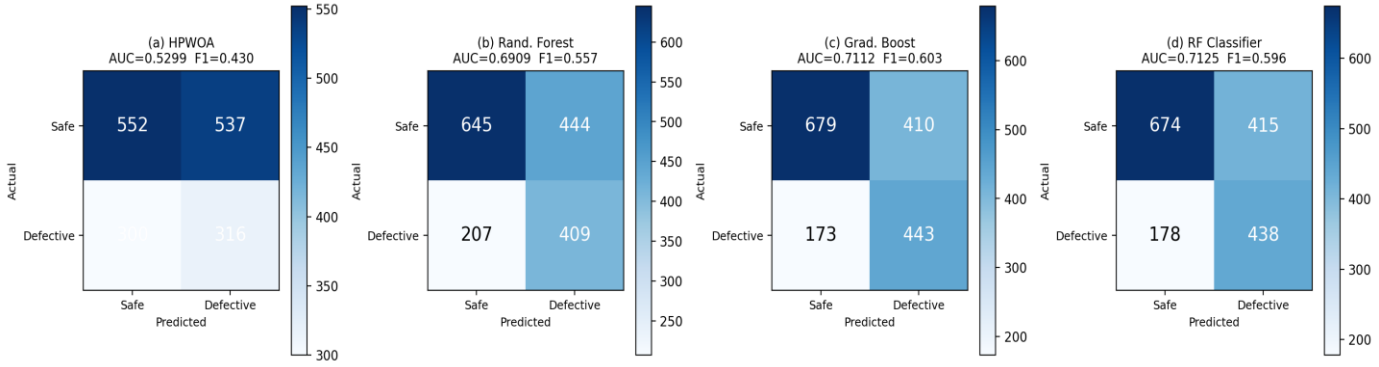


Figure 8. Confusion matrices at median risk score threshold.

8.5. Threshold Metrics and Top-Quartile Risk Capture

Table 5 presents the threshold-based classification performance and top-quartile risk-capture results for all evaluated models. At the median-risk threshold, GBM-Regressor achieves the highest sensitivity (0.719) and F1 score (0.603), while RF Classifier provides closely comparable performance with a sensitivity of 0.711 and an F1 score of 0.596. RF Classifier achieves the highest top-quartile risk capture, recovering 43.5% of all defective

windows within the highest-risk 25% of observations, followed by GBM-Regressor at 40.7%. Both exceed the corrected random baseline of 25.0%, demonstrating that the model outputs contain useful information for prioritizing potentially defective observations. HPWOA achieves a lower AUC and F1 score than the leading tree-based approaches but still produces a top-quartile capture of 34.6%, corresponding to a 38% improvement over random ranking. These results reinforce the finding that the principal value of the early-ramp models lies in relative defect-risk prioritization rather than precise estimation of absolute PDIV.

Table 5. Threshold-based classification metrics. * Highest F1; † highest AUC and TopQ.

Model	Sens.	Spec.	Prec.	F1	TP/FP/FN/TN	TopQ (rand = 0.250)	AUC [95% CI]
HPWOA	0.513	0.507	0.370	0.430	316/537/300/552	0.346 (+38%)	0.527 [0.389, 0.662]
Lin. Reg.	0.458	0.476	0.331	0.384	282/571/334/518	0.182 (−27%)	0.441 [0.322, 0.567]
RF-Regressor	0.664	0.592	0.479	0.557	409/444/207/645	0.377 (+51%)	0.693 [0.615, 0.770]
GBM-Regressor *	0.719	0.624	0.519	0.603	443/410/173/679	0.407 (+63%)	0.712 [0.624, 0.801]
Log. Reg. Cls.	0.627	0.571	0.458	0.526	386/489/230/600	0.354 (+42%)	0.627 [0.505, 0.752]
RF Classifier †	0.711	0.619	0.513	0.596	438/415/178/674	0.435 (+74%)	0.714 [0.628, 0.803]
Random baseline	≈0.361	≈0.639	-	-	-	0.250 (proof Section 7.3)	0.500

8.6. Feature Importance Analysis

Figure 9 compares the sensitivity, specificity, precision, and F1 score across the evaluated models. The results indicate that GBM-Regressor and RF Classifier provide the most favorable overall balance between the four classification measures. GBM-Regressor achieves the highest sensitivity of 0.719 and the highest F1 score of 0.603, while RF Classifier achieves a sensitivity of 0.711 and an F1 score of 0.596. RF-Regressor also performs competitively, with an F1 score of 0.557. HPWOA produces relatively balanced sensitivity and specificity values of 0.513 and 0.507, respectively, but its lower precision and F1 score indicate weaker overall classification effectiveness. These findings demonstrate that model performance should not be assessed using a single metric, as the ability to identify defective windows must be considered alongside false-positive control and overall classification balance.

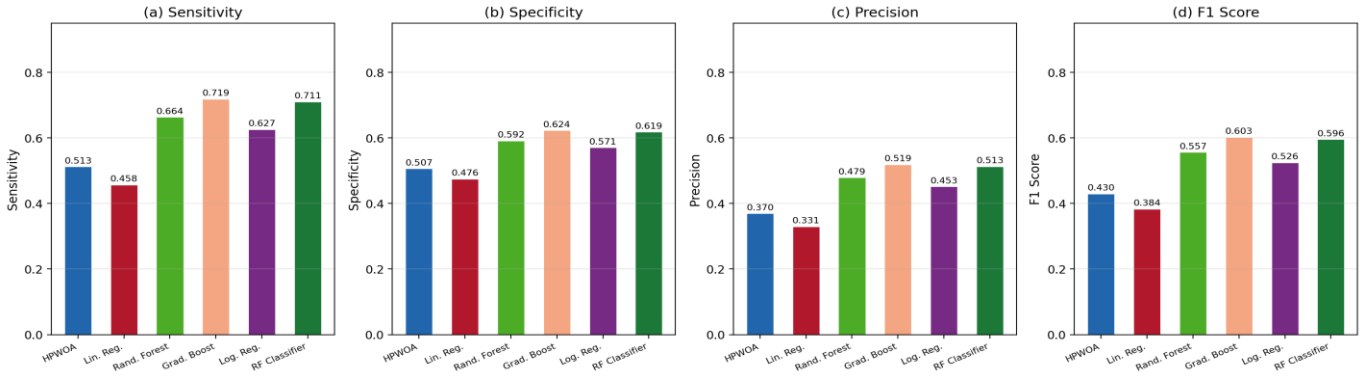


Figure 9. (**a**) Sensitivity, (**b**) specificity, (**c**) precision, and (**d**) F1 score.

Figure 10 provides two complementary perspectives on the practical usefulness and interpretability of the proposed early-risk framework. Figure 10a compares the top-quartile risk capture achieved by each model against the corrected random baseline of 0.250. RF Classifier achieves the highest top-quartile capture at 0.435, followed by GBM-Regressor at 0.407. These values correspond to improvements of approximately 74% and 63%, respectively, over random selection. HPWOA achieves a top-quartile capture of 0.346, corresponding to a more modest 38% improvement over random ranking. Thus, although the regression models do not accurately estimate PDIV itself, their continuous risk scores can nevertheless provide useful prioritization of potentially defective windows.

Figure 10b presents the feature-importance analysis for HPWOA and provides insight into why accurate PDIV regression remains difficult. GuarPD_kV is the dominant feature, indicating that the neural network relies strongly on the guaranteed test voltage, which is a transformer-level specification rather than a direct measurement of the early-ramp defect signature. MeanPD and Slope_Early contribute additional information related to discharge magnitude and its early temporal progression, but their contribution is insufficient to recover the transformer-level PDIV reliably from a 120 s window. The feature-importance results therefore support the interpretation that the model primarily learns voltage-class information while only partially capturing the early-ramp characteristics associated with insulation defects.

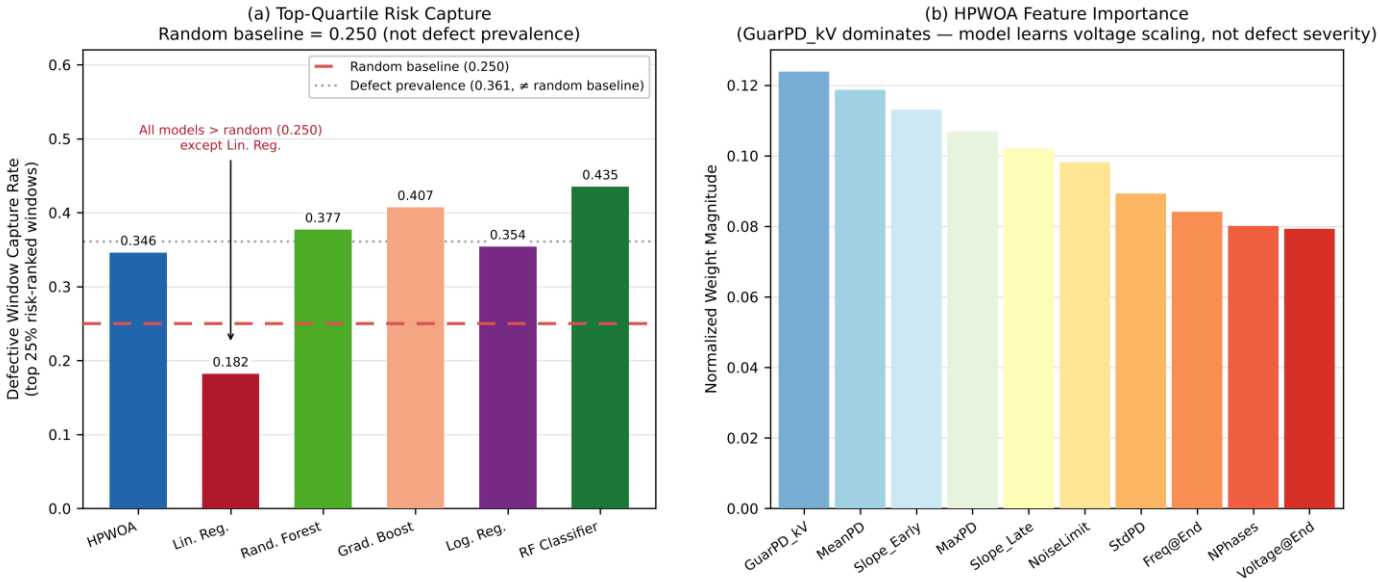


Figure 10. (**a**) Top-quartile risk capture vs. corrected random baseline, and (**b**) HPWOA feature importance.

9. Discussion

9.1. Why Regression R^2 Is near Zero

The near-zero and negative PDIV regression R^2 values are the empirical result of the reported file-grouped evaluation, not a theoretical requirement for every early-ramp dataset. PDIV depends on the insulation system, applied-voltage conditions, defect initiation, and test-specific measurement criteria; the available early-ramp descriptors evidently provide insufficient information for accurate absolute PDIV reconstruction. Figure 10b indicates strong dependence on the guaranteed test voltage, a voltage-class specification, while MeanPD and Slope_Early add more limited information about early discharge activity. For GBM-Regressor, PDIVRatio $R^2 = 0.097$ with the reported 95% file-bootstrap interval [−0.237, 0.354], which includes zero. This estimate is based on 39 withheld source

recordings from 18 physical transformers. Because some units contribute more than one file, the bootstrap interval does not quantify independence across transformers, and external out-of-transformer regression performance remains unknown.

9.2. Risk Ranking Value Despite Near-Zero R^2

For the recorded windows, GBM-Regressor's top-quartile capture of 0.407 means that the 25% highest-ranked windows contain 40.7% of all defective windows, compared with the random-selection expectation of 25.0%; RF Classifier increases the capture to 0.435. These are useful discrimination results for within-corpus, out-of-file triage, not established improvements in transformer-level failure probability or a prospectively validated safety intervention. HPWOA has a lower capture of 0.346. The ensemble methods outperformed the optimized neural network on the present small, heterogeneous tabular dataset, but this observation does not establish universal algorithm superiority. Dependence between repeated files from the same physical transformer must be evaluated before extrapolating those capture rates to new equipment.

9.3. Practical Deployment

A feasible future application is to calculate a risk score as each 120 s partial-discharge window becomes available on the acquisition workstation and to display it alongside the measured PD-magnitude trace. The retrospective study, however, evaluates windows selected from the first 50% of completed recordings and has not validated first-window-only decisions, real-time scoring latency, or prospective intervention effectiveness. A high score would prompt the test engineer to review the PD-magnitude time history, noise conditions, calibration, and the applicable standard test plan; it is not an instruction to terminate, extend, or otherwise modify a required IEC acceptance test. Before deployment, a proposed operating threshold must be validated using physically independent transformers and the prescribed safety procedures [8,36].

The generalizability is limited by the sample structure and collection setting. The available 39 test files represent only 18 transformers supplied by one utility and manufacturer within a single region. The present five-fold file-grouped evaluation controls overlap within each test recording but does not guarantee that the same physical transformer is absent from training when another of its recordings is tested. The 1000-replicate file-level bootstrap may also understate uncertainty at the transformer population level. Consequently, the AUCs and top-quartile capture rates reported here are retrospective recording-level estimates, not validated probabilities for new transformers, manufacturers, or sites. Future work should use a transformer-disjoint split, a two-level or equipment-cluster bootstrap, explicit 25–100% cutoff testing with voltage-stage identification, and independent multi-site datasets. The actual operational benefit will additionally depend on the facility's transformer-level defect prevalence rather than the window-level defect fraction in the present corpus.

10. Conclusions

The results support three conclusions. First, using information limited to the first 50% of a recorded IEC 60270 test permits retrospective risk ranking with useful discrimination, although the selection of 50% is a pragmatic trade-off rather than a standard-mandated or proven optimal decision point. The reported duration sensitivity increases from RF Classifier AUC 0.632 at 25% to 0.714 at 50%, 0.741 at 75%, and 0.758 at 100%. Second, early-ramp features do not accurately reconstruct absolute PDIV in the file-grouped experiment, whereas GBM-Regressor and RF Classifier achieve risk-ranking AUCs of 0.712 and 0.714 and top-quartile captures of 0.407 and 0.435, respectively. Third, the 39-file GroupK-Fold and file-bootstrap procedures address dependence among windows from the same

recording but cannot by themselves establish independence among separate test recordings from the same 1 of 18 physical transformers. These important limitations temper confidence intervals and claims of generalization to new transformers. No measured results have been recomputed in this revision.

HPWOA exhibits weaker discrimination in the present experiment (AUC = 0.527, 95% file-bootstrap interval [0.389, 0.662], top-quartile capture = 0.346) than the tree-based comparators. The substantive conclusion is therefore that the observed early signals contain some ranking information but limited absolute PDIV information. Prospective transformer-disjoint validation, equipment-cluster uncertainty estimation, and operating-stage-based sensitivity assessment are prerequisites for interpreting the reported ranking as a reliable equipment-level decision aid.

The present study is distinct from, and explicitly references, the authors' published 2025–2026 studies [24–28]. Those papers address frequency-response-based PD diagnosis, source or severity classification, or prediction of instantaneous PD magnitude from different feature and target definitions. This paper instead evaluates restricted early-window information for subsequent file-level PDIV and defect-risk prioritization. The earlier papers' numerical performance estimates are not used as matched baselines for this different endpoint.

Author Contributions: Conceptualization, L.T. and B.A.T.; methodology, L.T. and B.A.T.; software, L.T. and B.A.T.; validation, B.A.T.; formal analysis, L.T. and B.A.T.; investigation, L.T. and B.A.T.; data curation, L.T.; writing-original draft preparation, L.T.; writing-review and editing, L.T. and B.A.T.; visualization, L.T.; supervision, B.A.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The raw data used for model development and classifier training are not publicly available due to confidentiality restrictions imposed by the equipment owner. The code and simulation results supporting the findings of this study are available from the corresponding author upon reasonable request.

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

Abbreviations

The following abbreviations are used in this manuscript:

AUC	Area Under the ROC Curve
FNN	Feedforward Neural Network
GTD	Guaranteed Test Voltage
GWO	Grey Wolf Optimization
HPWOA	Hybrid Particle Whale Optimization Algorithm
HPWOA-DO-FNN	Hybrid Particle Whale Optimization Algorithm Dual-Output Feed Forward Neural Network
HV	High Voltage
IEC	International Electrotechnical Commission
PD	Partial Discharge
PDIV	Partial Discharge Inception Voltage
PSO	Particle Swarm Optimization
RF	Random Forest
ROC	Receiver Operating Characteristic
WOA	Whale Optimization Algorithm

References

1. Baul, S.; Dey, A.; Dey, J.; Dey, D.; Deb, S. Electricity Transmission Infrastructure: Engineering Principles, Economic Implications, and its Expanding Role in Renewable Energy. *ResearchGate* **2026**, https://doi.org/10.13140/RG.2.2.32043.60963.
2. Balanta, J.Z.; Rivera, S.; Romero, A.; Coria, G. Planning and Optimizing the Replacement Strategies of Power Transformers: Literature Review. *Energies* **2023**, *16*, 4448. https://doi.org/10.3390/en16114448.
3. CIGRE Working Group A2.37. *Transformer Reliability Survey*; CIGRE Technical Brochure 642; CIGRE: Paris, France, 2015.
4. Thiviyanathan, V.; Ker, P.; Leong, Y.; Abdullah, F.; Ismail, A.; Jamaludin, Z. Power Transformer Insulation system: A Review on the reactions, Fault Detection, Challenge and Future Prospects. *Alex. Eng. J.* **2022**, *61*, 7697–7713.
5. Mewally, I. Failures, Monitoring and New Trends of Power Transformers. *IEEE Potentials* **2011**, *30*, 36–43.
6. Adekunle, A.; Oparanti, S. Decay Products and Contaminants in Transformer Insulation: Mechanisms, Dielectric Impacts, and Sustainable Regeneration Strategies. *Next Sustain.* **2026**, *7*, 100305. https://doi.org/10.1016/j.nxsust.2026.100305.
7. Obiora, G.; Amoo, Y.; Igbinos, G.; Fiemobebefa, C.; Abamido, O. Transformer Failure Rates and Reliability Analysis. *Niger. J. Appl. Sci. Innov. Technol.* **2025**, *1*, 35–44.
8. *IEC 60270:2025*; High-Voltage Test Techniques—Charge-Based Measurement of Partial Discharge. International Electrotechnical Commission: Geneva, Switzerland, 2025.
9. Lu, S.; Chai, H.; Sahoo, A.; Phung, B. Condition Monitoring Based on Partial Discharge Diagnostics Using Machine Learning Methods: A Comprehensive State-of-the-Art Review. *IEEE Trans. Dielectr. Electr. Insul.* **2020**, *27*, 1861–1888.
10. Barrios, S.; Buldain, D.; Comech, M.; Gilbert, I.; Orue, I. Partial Discharge Classification Using Deep Learning Methods: Survey of Recent Progress. *Energies* **2019**, *12*, 2485. https://doi.org/10.3390/en12132485.
11. Khodaveisi, F.; Karami, H.; Karimpour, M.; Rubistein, M.; Rachidi, F. Partial Discharge Localization in Power Transformer Tanks Using Machine Learning Methods. *Sci. Rep.* **2024**, *14*, 11785. https://doi.org/10.1038/s41598-024-62527-9.
12. Ogliari, E.; Sakwa, M.; Wei, J.; Liu, W.; Schubert, B.; Palo, M. General Machine Learning-Based Approach to Pulse Classification for Separation of Partial Discharges and Interference. *IEEE Sens.* **2023**, *23*, 26839–26849.
13. Islam, M.; Chen, G.; Jin, S. An Overview of Neural Network. *Am. J. Neural Netw. Appl.* **2019**, *5*, 7–11.
14. Tomar, V.; Bansal, M.; Singh, P. Metaheuristic Algorithms for Optimization: A Brief Review. *Eng. Proc.* **2023**, *59*, 238. https://doi.org/10.3390/engproc2023059238.
15. Dokeroglu, T.; Sevinc, E.; Kucukyilmaz, T.; Cosar, A. A Survey on New Generation Metaheuristic Algorithms. *Comput. Ind. Eng.* **2019**, *139*, 106040.
16. Wang, H.; Xie, G.; Zhang, Y. Optimal Power Flow of Power Systems Using Hybrid Whale Particle Swarm Optimization Technique. In Proceedings of the IEEE International Conference on Energy Interney (ICEI), Zhuhai, China, 1–3 November 2024; pp. 598–602. https://doi.org/10.1109/ICEI63732.2024.10917184.
17. Bansaol, S.; Aggarwal, H. A Hybrid Particle Whale Optimization Algorithm with Application to Workflow Scheduling in Cloud-Fog Environment. *Decis. Anal. J.* **2023**, *9*, 100361.
18. Shami, T.; El-Saleh, A.; Alswaitti, M.; Al-Tashi, Q.; Mirjalili, S. Particle Swarm Optimization: A Comprehensive Survey. *IEEE Access* **2022**, *10*, 10031–10061.
19. Abualigah, L.; Sheikhan, A.; Ikotun, A.; Zitar, R.; Alsoud, A.; Al-Shourbaji, I.; Hussein, A.; Jia, H. Chapter 1—Particle Swarm Optimization Algorithm: Review and Applications. In *Metaheuristic Optimization Algorithms*; Abualigah, L., Ed.; Morgan Kaufmann: Burlington, MA, USA, 2024; pp. 1–14.
20. Nadimi-Shahraki, M.; Zamani, H.; Varzaneh, Z.; Mirjalili, S. A Systematic Review of the Whale Optimization Algorithm: Theoretical Foundation, Improvements, and Hybridizations. *Arch. Comput. Methods Eng.* **2023**, *30*, 4113–4159.
21. Liang, Z.; Shu, T.; Ding, Z. A Novel Improved Whale Optimization Algorithm for Global Optimization and Engineering Applications. *Mathematics* **2024**, *12*, 636.
22. Li, J.; Ersotelos, N.; Tsompanas, M.; Epiphaniou, G. Challenges and Opportunities of Synthetic Data Generation for Machine Learning-Based Intrusion Detection Systems in In-Vehicle Networks. *Veh. Commun.* **2026**, *58*, 100998. https://doi.org/10.1016/j.vehcom.2025.100998.
23. Pe, S.; Dagliati, A.; Parimbelli, E.; Guerra, B.; Sozzi, S.; Bellazzi, R.; Nicora, G. How Improper Dataset Split Hinders Model Generalizability: A Systematic Comparison in Human Activity Recognition and Exercise Evaluation Tasks. *Int. J. Med. Inform.* **2026**, *216*, 106464. https://doi.org/10.1016/j.ijmedinf.2026.106464.
24. Thango, B.A. Interpretation of Partial-Discharge-Activated Frequency Response Analysis for Transformer Diagnostics. *Machines* **2025**, *13*, 300. https://doi.org/10.3390/machines13040300.

25. Thobejane, L.T.; Thango, B.A. Partial Discharge Source Classification Using Machine Learning Algorithms. In *Proceedings of the 33rd Southern African Universities Power Engineering Conference (SAUPEC 2025), Pretoria, South Africa, 29–30 January 2025*; IEEE: New York, NY, USA, 2025. https://doi.org/10.1109/SAUPEC65723.2025.10944441.
26. Thobejane, L.; Thango, B.A. Linear and Nonlinear Feature Extraction for Transformer Partial Discharge Severity Classification: A Comparative Study Using Artificial Neural Networks. *Energies* **2026**, *19*, 2642. https://doi.org/10.3390/en19112642.
27. Thobejane, L.; Thango, B.A. Partial Discharge Severity Classification for Transformer Condition Monitoring Using Feature Engineering, PCA, and ANN. *Machines* **2026**, *14*, 711. https://doi.org/10.3390/machines14060711.
28. Thobejane, L.; Thango, B. Hybrid PSO-WOA-Optimized Neural Network for Partial Discharge Prediction in Power Transformer Acceptance Testing. *Algorithms* **2026**, *19*, 621. https://doi.org/10.3390/a19080621.
29. Basharan, V.; Rajamayil, M. Defect Diagnosis in Power Transformer Using Explainable AI: A Semi-Supervised Learning-Based Partial Discharge Pattern Approach. *Electr. Power Syst. Res.* **2025**, *247*, 111860. https://doi.org/10.1016/j.epsr.2025.111860.
30. Zhang, J. A Multi-Task Learning Framework for Diagnosing Partial Discharge Types and Assessing Severity. *Sci. Rep.* **2026**, *16*, 26885. https://doi.org/10.1038/s41598-026-57520-3.
31. Owczarczuk, B.; Dziadak, B.; Starzyński, J. Embedded AI/ML Systems for Partial Discharge Monitoring: A Review. *Energies* **2026**, *19*, 4411. https://doi.org/10.3390/en19184411.
32. Eckert, M.; Mraz, P.; Ballweg, C.; Pihera, J. Partial Discharge Behavior and Insulation Failures Detection in Electrical Devices Subjected to Impulse Voltage Excitation. *Int. J. Electr. Power Energy Syst.* **2025**, *172*, 111078.
33. Emdadi, K.; Gandomkar, M.; Aranizadeh, A.; Vahidi, B.; Mirmozaffari, M. Overview of Monitorinng, Diagnostics, Aging Analysis and Maintenance Strategies in High-Voltage AC/DC XLPE Cable Systems. *Sensors* **2025**, *25*, 70–96. https://doi.org/10.3390/s25227096.
34. Elorza Azpiazu, L.; Almandoz, G.; Egea, A.; Ugalde, G.; Badiola, X. Study of Partial Discharge Inception Voltage in Inverter Fed Electric Motor Insulation Systems. *Appl. Sci.* **2023**, *13*, 2417. https://doi.org/10.3390/app13042417.
35. Naderiallaf, H.; Giangrande, P.; Galea, M. Characterization of PDIV, PDEV and RPDOV in Insulated Wires Under Unipolar Repetitive Squar Wave Excitations for Inverter-Fed Motors. *IEEE Access* **2023**, *11*, 51047–51063. https://doi.org/10.1109/ACCESS.2023.3279886.
36. *IEC 60076-3:2013*; Power Transformers—Part 3: Insulation Levels, Dielectric Tests and External Clearances in Air. International Electrotechnical Commission: Geneva, Switzerland, 2013.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.