*Proceedings of the International Conference on Industrial Engineering and Operations Management*

Publisher: IEOM Society International, USA DOI: 10.46254/AP07.20260394
Published: March 25, 2026

# Integrated Machine Learning and Economic Risk Modeling for Reliability-Centered Asset Management in Electrical Distribution Networks

**Md. Efatuzzaman Efat**
Institute of Information and Communication Technology
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
efatuzzaman@gmail.com

**Irtefa Waseek**
Department of Statistics and Data Science
Jahangirnagar University
Dhaka, Bangladesh
waseekirtefa@gmail.com

**Sumiya Afrose**
Department of Robotics and Mechatronics Engineering
University of Dhaka
Dhaka, Bangladesh
s.afrose098@gmail.com

**Sayed Md. Fazle Rabbi**
Department of Electrical and Electronic Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
sayedfazlerabbi45@gmail.com

**Md. Saiful Islam**
Professor
Institute of Information and Communication Technology
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
mdsaifulislam@iict.buet.ac.bd

## Abstract

In electrical distribution systems, unforeseen transformer faults have caused service interruptions, deteriorated supply quality and economical damage on unprecedented scale over the past decades. Although predictive maintenance has adopted machine learning techniques, most existing studies have concentrated on classification accuracy without taking into proper account probability calibration, economic consequences modeling, or time resilience. This paper introduces an economic risk and machine learning model for the prediction of rare events in transformer burnout within distribution grids. One years of operational data are employed to train models using historical data and to test them on a future year to mimic real operating environments. A number of imbalanced learning strategies and ensemble algorithms are utilized and evaluated using precision-recall metrics, which account for the rarity of transformer failures. The study involves the use of probability calibration to make decisions more reliable and to maximize the benefits of risk management in terms of monetary value. Furthermore, the predictive results are attributed to engineering-relevant stress factors, such as outage exposure, network density, and environmental conditions, using explainable analysis. In addition to predictive performance, the expected failures are converted into expected monetary risk is provided by accounting for replacement costs, outage value, and preventive maintenance efficacy. A life-cycle economic model, including discounting, is applied to determine the optimal inspection portfolio to maximize net present value and support reliability-centered maintenance decision making. Sensitivity analysis and Monte Carlo simulations are used to indicate the robustness of the economic results in the face of parameter uncertainty.

## Keywords
Cost-Sensitive Learning, Economic Risk Modeling, Transformer Failure Prediction, Distribution Network Reliability and Smart Grid Asset Management.

## 1. Introduction
Predictive maintenance (PdM) has emerged as a data-driven strategic tool for the failure prediction in systems to enable corrective action prior to breakdowns. Current maintenance strategies are primarily dependent on reactive maintenance which is often performed after sudden outages, resulting in high-consequence damage and poor service. In this context, the deployment of PdM can transform these unplanned post-failure maintenance activities into pre-planned interventions. In distribution networks, transformers play an important role as nodes and its failure can cause customer interruptions to propagate and lead to direct and indirect economic losses. The economic dimension of transformer outages is commonly summarized by the value of lost load (VoLL), which is a tool for reliability decision-making as well as cost-benefit analysis (Gorman 2022). PdM is also a key enabler for smart grids, as intelligent operation requires monitoring, analytics, and proactive decision-making instead of post-event reactions. System-level reviews of PdM in smart-grid contexts are directed at the use of IoT/AMI measurement flows and analytics for proactive fault detection and resilience improvement (Mahmoud et al. 2021). From the practical point of view of power assets, PdM concepts are used for transformer fleets, switchgear, and feeders, where risk screening can help prioritize limited maintenance crews for the highest-impact assets (Vita et al. 2023).

### 1.1 Objectives
We organized this research in order to meet five practical objectives:
1. Temporal Generalization using one-year dataset as the training set and another one-year dataset the testing set to model deployment under a distribution shift rather than splitting randomly.
2. To deploy multiple algorithms with different learning strategies and benchmark using precision-recall metrics.
3. To redirect forecast error for probability calibration to enhance decision support.
4. To improve model interpretability by explaining predictions in terms of physically meaningful factors using SHAP-based attribution analysis.
5. To optimize portfolio performance for its cumulative net present value, taking into account reliability through Monte Carlo and sensitivity analysis.

## 2. Literature Review
Transformer predictive maintenance research covers the areas of (i) condition monitoring and fault diagnosis (often through dissolved gas analysis, DGA) and (ii) fleet-level failure prediction based on operational and contextual features. Reviews summarize ample applications of AI/ML to transformer fault diagnosis and show that challenges remain in data imbalance, generalization, and interpretability (Wani et al. 2021; Zhang et al. 2022).Recent diagnostic studies tend to adopt deep learning and explainability for better diagnostic robustness; for example, explanation-based architectures with SHAP explanations and explicit imbalance treatment can be found in recent transformer fault analysis work (Lin et al. 2025). At the distribution level, publicly available fleet datasets are still limited, which is in turn a limitation for reproducibility. An exception to this is the data on transformers in Cauca Department from Data in Brief, hosted Mendeley Data that was collected with the cooperation of the grid operator Compañía Energética de Occidente and Colombian universities (Bravo M et al. 2021) . Using this dataset family, a machine learning

methodology for the predictive maintenance of distribution transformers was proposed by Alvarez Quiñones et al. (Alvarez Quiñones et al. 2022), while Vita et al. (2023) examined predictive maintenance for DSOs to enhance transformer reliability. While these works demonstrate the feasibility of fleet prediction, they also serve as motivation for more robust pipelines for end-to-end prediction-decision linking probability outputs and economic decisions.

From a methodological perspective, much of the modern PdM literature exploits ensemble learners and gradient boosting, due to their excellent performance on tabular data. XGBoost (Chen and Guestrin 2016), LightGBM (Ke et al. 2017). These scalable boosters include boosted trees (Friedman 2001). At the same time, explainable ML has gone from good to essential with critical infrastructure; SHAP attributes are theoretically well-grounded, and are now widely used in the analytics of power-assets (Lundberg and Lee 2017).

Two persistent gaps still remain. First, many studies report classification accuracy but do not address probability calibration, especially when decisions are based on the quantification of risk. It is important that calibrated probabilities are obtained in classification tasks, and recent calibration studies pay more attention to the real impact of post-hoc scaling methods (Guo et al. 2017). Second, studies often stop at ranking or classification and do not make the most of economic consequence modeling. In reliability planning, outage cost is usually quantified in terms of VoLL (Value of Lost Load) and used in cost-benefit frameworks and regulatory reliability standards(Gorman 2022; ACER 2020). Risk-aware scheduling of transformer maintenance has been tackled using either expected-utility or cost-based frameworks (Sarajčev et al. 2020) but has rarely been integrated with calibrated machine learning (ML) probabilities, temporal holdout validation, and portfolio optimization under uncertainty within one unified pipeline.

**Our study fills these gaps by combining:**
(a) We employ temporal validation using a "train-on-past, test-on-future" design to ensure that predictive performance reflects realistic deployment conditions.
(b) We incorporate probability calibration, thereby producing decision-grade risk estimates suitable for cost-sensitive applications.
(c) We emphasize explainability for interpretability of model outputs by engineers and decision-making transparency.
(d) We also consider discounted life-cycle (NPV) economics and use this information to optimize the inspection portfolio, supported by sensitivity analyses and Monte Carlo runs integrating parameter uncertainty in order to increase confidence in our recommendations.

# 3. Methodology
## 3.1 Dataset description
We used data on distribution transformers from the Cauca Department in Colombia, which was published as an open-access article in the journal Data in Brief by Bravo et al. and is available on Mendeley Data (Bravo M et al. 2021). The dataset is derived from real-world operational records gathered through a joint initiative between a regional grid operator and several prominent academic institutions in South America. This partnership ensures the data is of both industrial and academic relevance. As represented in Table 1, each year includes 15,873 transformers and 15 binary base features, with "Burned transformers" as the binary target. Failures are rare—5.08% in 2019 (807/15,873) and 3.96% in 2020 (629/15,873)—which motivates the use of imbalanced learning evaluation.

Table 1. Dataset summary (temporal split: 2019→2020)

| Dataset Characteristics | Development Set (2019) | Evaluation Set (2020) |
|---|---|---|
| Total Population (N) | 15,873 | 15,873 |
| Failure Events (Minority) | 807 | 629 |
| Operational Units (Majority) | 15,066 | 15,244 |
| Failure Prevalence | 5.08% | 3.96% |
| Features Count | 15 | 15 |

**Engineering Interpretation of Dataset Variables:** The transformer reliability dataset described in the Data in Brief publication consists of several variables representing the operational, environmental, and network characteristics of distribution assets. The important characteristics are associated with the classification of the location (urban or rural), the rated power of the transformer (kVA), the arrangement of the protection, and the differences indicating the lightning activity (average and maximum lightning discharge density). In addition, the dataset contains the number of connected customers, the network length of this customer, the Expected Energy Not Supplied (EENS). All these variables depict physical capacity of the asset and the exposure of the network to environmental and operational stresses, which are major root causes of risk of transformer failure (Bravo M et al. 2021).

In Figure 1, the 2019 and 2020 datasets of class distributions of transformer failures were presented, with the division between regular work and failures being a substantial one. The figure also indicates that the prevalence of failure is also seen to be declining in 2020 than in the year 2019. This type of high-class imbalance is typical of reliability datasets and encourages the use of precision-recall evaluation measures and top-k inspection prioritization strategies instead of relying solely on accuracy. These are the evaluation options that are more compatible to the operational objective of labeling few of risky assets and providing a priority of maintenance on those assets.

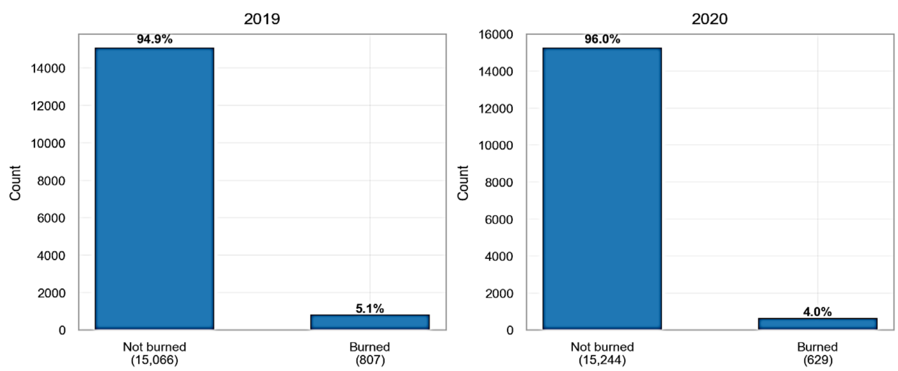


Figure 1. Class Distribution and Frequency of Failure across Different Years

### 3.2 Data Preparation and Predictive Model Development Pipeline

The predictive maintenance framework designed in this study follows a structured data-processing and modeling pipeline aimed at connecting transformer reliability prediction with economic decision support. The complete workflow is shown in Figure 2, which summarizes the sequential data preparation, model development, evaluation, and decision analysis.

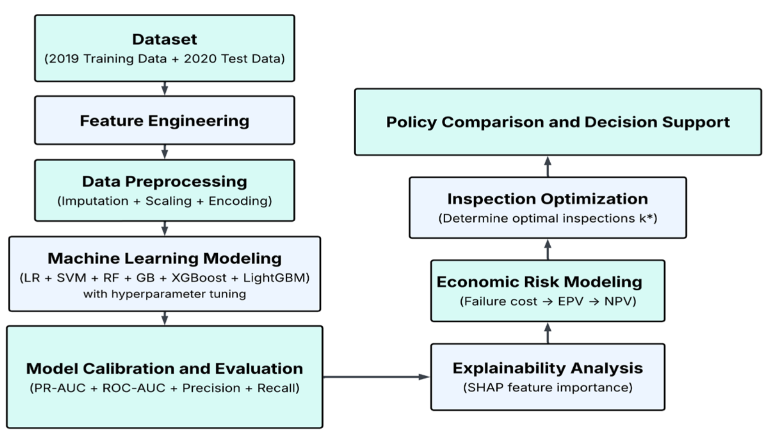


Figure 2. Predictive maintenance and economic decision support pipeline for transformer reliability management.

The process begins by assembling the transformer datasets from 2019 and 2020. We use the full 2019 records for model training, while setting aside the entire 2020 dataset exclusively for out-of-sample testing. Rather than relying on a random split, we deliberately adopt a temporal split to mirror real-world operating conditions, where predictive models must be trained on historical observations before being asked to forecast failures that lie ahead. This method also stops the leaking of time-related information and gives a more accurate picture of how well a model generalizes in reliability-centered asset management. After the datasets are combined, feature engineering is used to create important variables that show how the transformer works, how it is exposed to the environment, and how the network works that are important for its reliability. The subsequent stage involves data preprocessing, which prepares the feature set for machine learning analysis. This process encompasses the imputation of missing values, normalization of numerical variables, and encoding of categorical attributes such as location type and protection configuration. These transformations ensure that all predictors are expressed in a standardized format, thereby facilitating their effective utilization by machine learning algorithms. Following preprocessing, several supervised machine learning models are trained to assess the risk of transformer failure and support reliability-based maintenance planning. The models evaluated include Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), Gradient Boosting (GB), XGBoost, and LightGBM. Each algorithm has its own optimizable predictive performance hyperparameters. These scores represent the failure probability for a transformer's failure risk as modeled by our algorithms, and can be used as inputs to mark agency inspections or as indicated on various economic analyses of cases that arise from predictive models. At the calibration and evaluation stage, model performance is evaluated using metrics such as Precision-Recall Area Under the Curve (PR-AUC), Receiver Operating Characteristic Curve Area under the ROC Curve (ROC-AUC), precision and recall that are suitable for imbalanced reliability data sets. These metrics provide a low-bias assessment of model performance in detecting rare failure events. Beyond predicting, our framework brings together explanations and economic decision model. SHAP (SHapley Additive exPlanations) is used to interpret predicted failure risk. Next, the predicted probabilities are used in an economic risk modeling stage, converting failure costs into Expected Present Value (EPV) and Net Present Value (NPV) measures. Empirical data of these economic indicators help in performing the inspection optimization step which decides on the maximally profitable amount k of inspections to conduct. The outputs are then consolidated in a policy comparison and decision-support stage. This helps utilities assess different maintenance strategies and prioritize high-risk transformers. This integrates pipeline links machine learning-based failure prediction with reliability-centered asset management decisions in electrical distribution networks.

### 3.3 Model Assessment Matrices

We consider both quality and value in terms of predictive capability; whenever a maintenance model is proposed, it must be actionable in terms of operations and defensible in terms of financial value.

**Non-economic measures of performance:** Regarding predictive measurements that exclude economic factors, we adopt PR-AUC (average precision) as our primary model-selection metric under class imbalance, in addition to ROC-AUC for discrimination analysis. The relationship between ROC-AUC, PR-AUC, and imbalance is non-trivial; therefore, we have presented both without claiming that one "universally dominates" the other. To visualize how well our probabilities align with reality, we utilized calibration curves. An additional layer of performance assessment was implemented through the computation of precision and recall metrics, construction of confusion matrices, and establishment of a specific decision threshold to inform result interpretation. Threshold selection was performed by optimizing the F2-score (with beta set to 2) on the 2019 validation set, an approach that places greater emphasis on recall relative to precision. This weighting scheme aligns with the operational objective of early failure detection within a predictive maintenance framework.

**Economic measures of performance:** Based on calibrated annual failure probabilities, we calculate the NPV at the transformer level for the decision to take preventive action. This is based on a discounted multi-year progression using VoLL-based outage cost modeling. We report the cumulative NPV versus the number of inspections, the maximum NPV, and the economically optimal number of inspections (k*). We also report the cumulative avoided Expected Present Value (EPV) of failures and the cost per avoided failure, enabling a direct financial justification of the maintenance portfolio.

## 4. Result

### 4.1 Benchmarking Failure Prediction Capabilities

Failure prevalence in the deployment like 2020 test set is 3.96% as showed in table 1. Since failures are infrequent, using information about total accuracy can be deceptive. For instance, a naive "most frequency class" baseline predicting all the transformers as "not burned", will obtain an accuracy of about 0.960, but will detect zero failures (recall = 0). This illustrates why precision-recall-based evaluation is necessary when modeling rare failure events.

Table 2. A comparison of resampling methods for the baseline Logistic Regression model Interpretation

| Strategy | ROC-AUC | PR-AUC | F2 |
|---|---|---|---|
| No resampling | 0.796814 | 0.252385 | 0.392523 |
| SMOTE | 0.798168 | 0.250766 | 0.407341 |
| Class weights | 0.806277 | 0.242883 | 0.414176 |

(Table 2): Class weighting results in the best measure of ROC-AUC and F2 score in this controlled comparison and the No-resampling configuration results in the slightly higher PR-AUC. Because there is no single imbalance strategy that is clearly dominant across all the metrics, we have kept both class-weighting and algorithm-specific imbalance treatments in the broader (that is, model benchmark) because the approach to this is not the sole way of handling such situations.

Key finding (Table 3): Logistic Regression has the best PR-AUC (0.153) and is therefore the best in terms of overall performance in terms of precision-recall trade-off on the 2020 deployment-like test set. In contrast, Random Forest obtains the highest ROC-AUC (0.733) and the highest F2 score (0.303), which indicates a good & practical tradeoff between ranking and recall-weighted performance of classification.

Table 3. Results for the 2020 test set (Decision thresholds optimized using 2019 F2 validation).

| Model | ROC-AUC | PR-AUC | F2 | F1 | Accuracy | Precision | Recall | Lift@5% | Threshold | P@5% | R@5% |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Logistic Regression | 0.727 | 0.1531 | 0.282 | 0.165 | 0.783 | 0.0972 | 0.539 | 5.696 | 0.05 | 0.226 | 0.285 |
| SVM (RBF) | 0.725 | 0.1404 | 0.294 | 0.182 | 0.823 | 0.112 | 0.498 | 5.123 | 0.07 | 0.203 | 0.256 |
| XGBoost | 0.732 | 0.1343 | 0.291 | 0.208 | 0.880 | 0.141 | 0.397 | 4.742 | 0.13 | 0.188 | 0.237 |
| Random Forest | 0.733 | 0.129 | 0.303 | 0.209 | 0.870 | 0.137 | 0.434 | 5.378 | 0.08 | 0.213 | 0.269 |
| LightGBM | 0.720 | 0.105 | 0.271 | 0.183 | 0.859 | 0.119 | 0.399 | 4.0733 | 0.1 | 0.161 | 0.2035 |
| Gradient Boosting | 0.7045 | 0.092 | 0.262 | 0.186 | 0.876 | 0.126 | 0.3593 | 3.405 | 0.1 | 0.135 | 0.1701 |

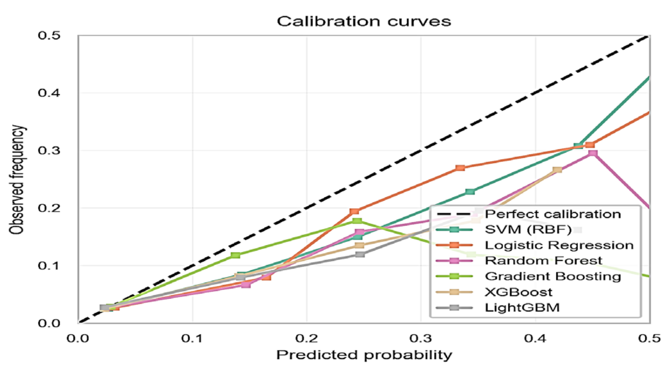


Figure 3. Calibration curves following sigmoid calibration.

Figure 3 presents calibration curves that compare predicted probabilities against observed failure frequencies. The diagonal line denotes perfect calibration, while deviations from it signal systematic bias in probability estimation. Since our framework directly incorporates predicted failure probabilities into the economic layer, calibration is not a superficial adjustment—it materially influences the monetization of risk and the prioritization of inspection portfolios.

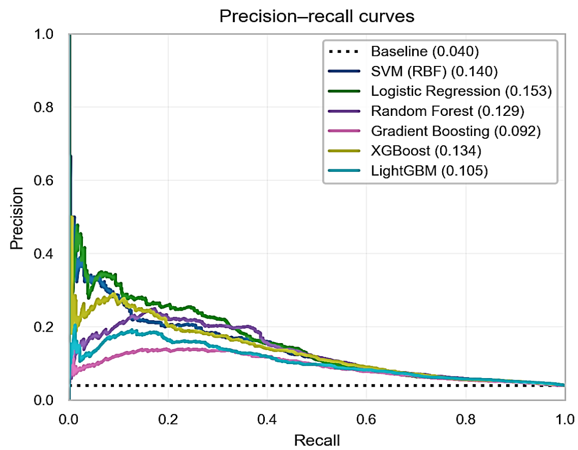


Figure 4. Precision-Recall curves on calibrated models.

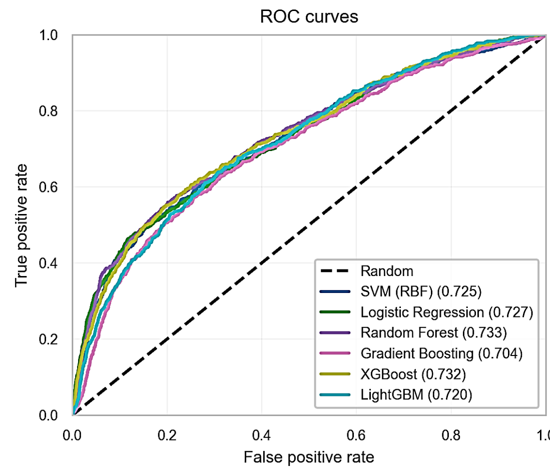


Figure 5. ROC curves on calibrated models.

The precision-recall curve shows how the precision and recall change with different thresholds. The dotted line in Figure 4 shows how often things went wrong in 2020 (0.040). Any model curve that is above this baseline will show that it has a useful ability to rank predictions. Logistic Regression gets the highest PR-AUC, which is in line with what Table 3 says and suggests that it should be the main tool for ranking probabilities in terms of risk monetisation. Furthermore, ROC curves describe the discrimination ability of the model for all possible classification thresholds. As shown in Figure 5, the majority of models cluster tightly within a ROC-AUC range of approximately 0.70–0.73, indicating broadly comparable discriminative performance. We therefore evaluate and interpret results using both ROC-AUC and PR-AUC. While ROC-AUC remains relatively robust in the presence of class imbalance, PR-AUC is highly sensitive to the prevalence of the positive class and thus provides a more informative measure for rare-event detection problems such as transformer failure prediction.

To evaluate the robustness of the economic optimization, we conducted sensitivity analysis and Monte Carlo simulations. The sensitivity analysis varies key economic parameters such as the value of lost load (VoLL) and preventive effectiveness, while the Monte Carlo simulation samples uncertainty distributions to assess the stability of the optimal inspection portfolio.

## 4.2 Failure Detection Effectiveness for Inspection Planning

In actual distribution system operation, utilities rarely inspect every asset in the system. Instead, inspections are usually focused on a small population of high-risk transformers. To reflect this operational reality, we report P@5%, R@5%, and Lift@5% (Table 3). Logistic Regression achieves a Lift@5% = 5.70, with 28.46% of all failures being captured by inspecting only the top 5% of ranked transformers (R@5% = 0.2846) with a precision of P@5% = 0.2257. SVM and Random Forest show similar results in our measure of top-k evaluation, while Gradient Boosting is evidently worse for our prioritization measures.

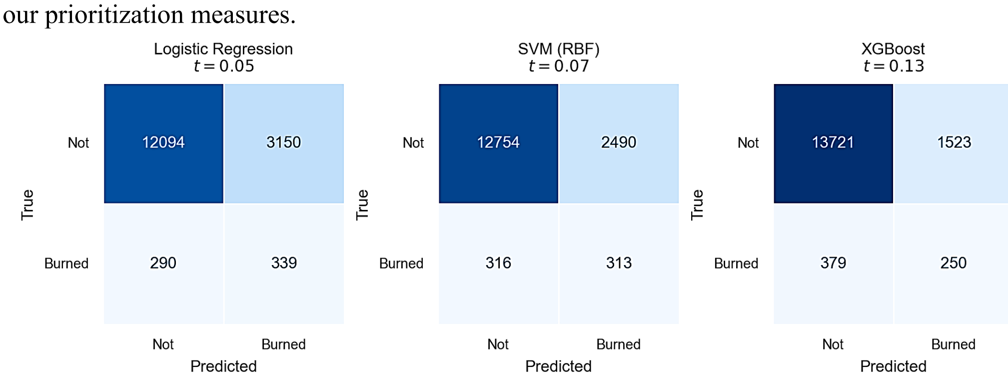


Figure 6. Confusion matrices for top-3 PR-AUC models: Validation-tuned thresholds.

These confusion matrices in Figure 6 are the operational decision threshold in which F2 is maximized on the 2019 validation set and then 2020 test data is fed through the model. As an example, events detected by Logistic Regression (threshold of about 0.05) are 339 out of 629 burned transformers (recall of about 0.539), but generates 3,150 false positives. This shows that there is an aggressive screening policy to give precedence to recall. XGBoost, conversely, uses a more conservative threshold, thus yielding a smaller number of false alarms but poorer recall. This comparison indicates the direct influence of the threshold selection on the operational maintenance effort.

### 4.3 Evaluating the Physical Drivers of Model Predictions

To model engineering reasoning within the system, we apply SHAP explanations to the top-performing tree model from the benchmark (XGBoost) using the TreeExplainer, consistent with established SHAP methodology (Lundberg and Lee 2017).

The SHAP summary (Figure 7) reveals that exposure to outages and network stress characteristics are prominent drivers of risk. Specifically, EENS (including engineered features like EENS per km/user) features strongly; this indicates that higher outage exposure is directly associated with an elevated risk of burnout. Additionally, power ratings and power-per-user emerge as significant factors, which is consistent with capacity-driven thermal and electrical loading intensity. Environmental factors such as lightning density and discharge ratios are also significant, correlating with the known vulnerability of distribution assets to external electrical stressors.

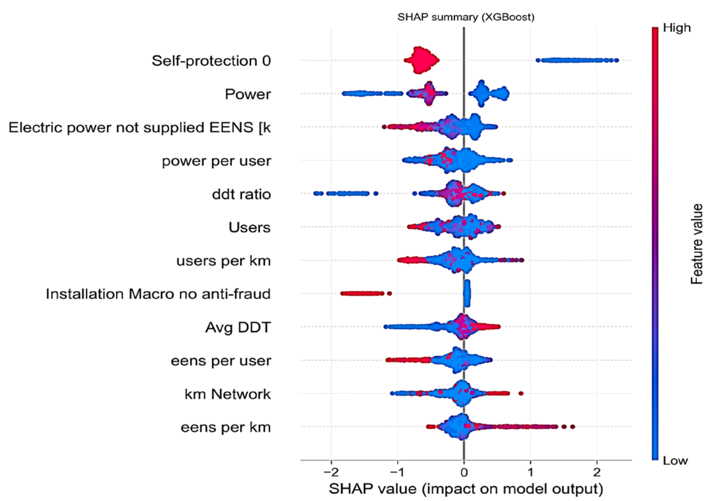


Figure 7. SHAP summary plot of XGBoost (top predictors).

### 4.4 Economic Optimization of Preventive Maintenance

For each transformer, we calculate a failure consequence cost which consists of:

1. Replacement cost proxy scaled by the transformer power rating.
2. Outage cost in proportion to EENS multiplied by VoLL.
3. Emergency crew cost.

VoLL-based outage valuation is standard practice in reliability economics and has wide use in cost-benefit analysis and reliability standard development (Gorman 2022; ACER 2020). We then estimate the expected present value (EPV) of failure over a multi-year time period using discounting, and calculate:

$$NPV = avoided\ EPV - program\ cost$$

where the preventive inspection and intervention are applied.

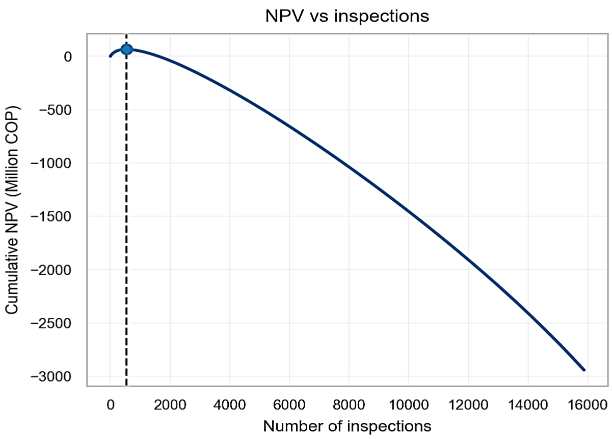


Figure 8. Cumulative NPV vs number of inspections.

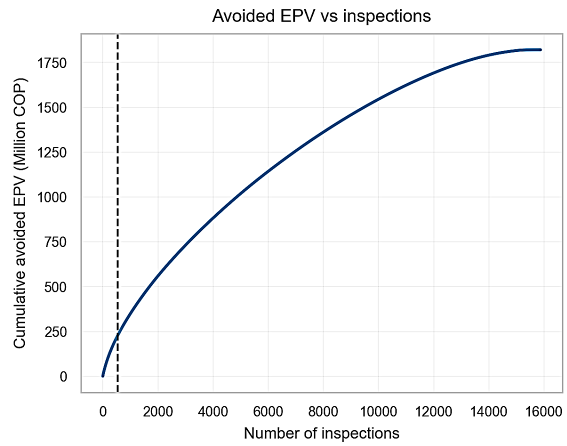


Figure 9. Cumulative avoided EPV versus number of inspections.

The cumulative NPV curve (Figure 8) reaches a maximum at the economically optimal number of inspections, k* (represented by the vertical dashed line). Consequently, inspecting beyond this point results in a loss, since the value of the avoided risk from further inspections decreases and eventually becomes less than the cost of the inspection and preventive actions. This result clearly shows that more maintenance does not necessarily create more value. In a similar vein, Figure 9 shows that avoided EPV increases as more transformers are inspected, but the rate of increase gradually slows down due to diminishing returns. When interpreted together with Figure 8, this behavior illustrates the key economic mechanism: averted risk eventually saturates while program costs continue to accumulate, thereby producing an interior optimum rather than a strategy of checking all assets.

## 4.5 Economic Policy Comparison

We compare four ways of making practical inspection policies using the same inspection budget size (k = k* for top-k policies): ranking by NPV (our method), ranking by probability only, random inspection and inspecting all.

Interpretation (Table 4): Ranking by NPV provides both the highest NPV and the highest ROI, which suggests that monetized risk ranking outperforms probability-only ranking when costs and consequences vary across assets. In contrast, while inspecting all assets maximizes the number of failures caught (recall = 1), it yields a severe negative NPV due to the excessive program costs. This comparison formally demonstrates the necessity of economic optimization to balance reliability benefits against financial realities and budget constraints.

Table 4. Performance of different inspection policies on the 2020 transformer fleet (k = 533).

| Inspection Policy | k | NPV (COP) | ROI | Avoided EPV (COP) | Program Cost (COP) | Failures Detected | Recall | EENS Prevented (kWh, 1y) | Cost per Avoided Failure (COP) |
|---|---|---|---|---|---|---|---|---|---|
| Top-k by NPV | 533 | 65425000 | 0.4092 | 225325000 | 159900000 | 112 | 0.1780 | 30672.1 | 1427680 |
| Top-k by probability | 533 | 6415060 | 0.0401 | 166315000 | 159900000 | 134 | 0.213 | 26414.7 | 1193280 |
| Random k | 533 | -98021000 | -0.613 | 61879000 | 159900000 | 23 | 0.0366 | 5345.18 | 6952170 |
| Inspect all | 15873 | -2941220000 | -0.618 | 1820680000 | 4761900000 | 629 | 1 | 159795 | 7570590 |

# 5. Discussion

In Table 5, we situate the present study within recent literature addressing reliability modeling and predictive maintenance of transformers. Many previous studies have contributed significantly to the establishment of the dataset, machine learning prediction and diagnostic interpretability. Few works formally couple calibrated predictive models with discounted economic optimization and temporally validated decision pipelines. Therefore, the ongoing conversation considers predictive performance and techno-economic viewpoints to highlight the contribution of predictive models in taking actionable maintenance decisions.

Table 5. Comparison with recent transformer reliability and predictive maintenance studies

| Study | Asset / Data | ML approach | Calibration for decision-grade probabilities | Economic optimization (NPV/utility) | Temporal validation (train past → test future) | Explainability |
|---|---|---|---|---|---|---|
| Bravo M et al. (2021) | Distribution transformer fleet dataset (Cauca, Colombia) | Data release enabling ML research | Not the focus | Not the focus | Provides multi-year dataset | Not the focus |
| Alvarez Quiñones et al. (2022) | Distribution transformers (Cauca) | ML for predictive maintenance scheduling | Not emphasized | Scheduling-oriented, limited explicit NPV portfolio | Uses the dataset context | Limited |
| Vita et al. (2023) | Distribution transformers (Cauca) | ML-based failure prediction for DSOs | Not central | Limited explicit discounted NPV portfolio | Year-based evaluation context | Some |
| Richardson et al. (2024) | General ML evaluation | Metric analysis for imbalanced classification (ROC vs PR) | Not applicable | Not applicable | Not applicable | Not applicable |
| Lin et al. (2025) | Power transformer fault diagnosis (DGA) | GAN + deep learning + SHAP | Notes importance of reliability/calibration | Not portfolio NPV-focused | Not fleet-level temporal validation | Strong |
| This study | Distribution transformer fleet dataset (Cauca, Colombia) | Multiple ML models (LR, SVM, RF, GB, XGBoost, LightGBM) for failure risk ranking | Discusses probability calibration for economic risk estimation | Discounted NPV-based inspection portfolio optimization with optimal k* | Strict temporal validation (train 2019 → test 2020) | SHAP-based interpretation of engineering risk drivers |

Our findings highlight that reliability-centered asset management should not be confined to the task of “failure prediction.” Instead, it must extend toward the broader objective of optimizing interventions under conditions of cost constraints and uncertainty.

First, Figures 4–5 and Table 3 demonstrate a moderate level of discriminatory performance (ROC-AUC approximately 0.70–0.73), yet they also reveal more pronounced variation in the ranking of infrequent events (PR-AUC approximately 0.09–0.15). Among the evaluated models, Logistic Regression achieves the highest PR-AUC, suggesting its relative strength in prioritizing rare failures. In contrast, Random Forest yields the strongest F2-score at the validated threshold, indicating that it may be more appropriate in contexts where recall-weighted classification is prioritized over risk-based ranking.

Second, Figure 3 underscores the importance of probability calibration. Because risk is quantified in terms of EPV/NPV, any systematic bias in predicted probabilities directly distorts expected cost estimates, potentially leading to mis-ordered inspection priorities. This observation aligns with broader evidence in the calibration literature, which shows that post-hoc scaling can substantially improve decision quality when probability estimates are used as quantitative inputs (Guo et al. 2017).

The techno-economic insights derived from Figures 8–9 and Table 4 highlight a pivotal tension in asset management: while the value of avoided risk eventually plateaus, program expenditures scale linearly. This divergence indicates that an exhaustive "blanket" inspection strategy is economically sub-optimal. Instead, the data supports an optimal inspection threshold (k) that maximizes utility. Our findings demonstrate that an NPV-optimal policy yields a positive net present value and a robust ROI, whereas both random sampling and universal inspections are economically dominated. A critical observation is that selection based strictly on failure probability—while effective at identifying more total failures for a given k—results in a markedly lower NPV. This confirms that failure probability is not a direct proxy for economic value when consequences, such as Expected Energy Not Served (EENS) and localized replacement costs, vary significantly across the asset base.

## 6.Conclusions and Future Work

This investigation proposed a reliability-focused framework to achieve operational deployment, designed to predict distribution transformer burnout and inform maintenance decisions. The approach was validated using a strict temporal holdout, training on historical data from 2019 and testing on observations from 2020, employing an authentic transformer dataset obtained from a utility partner. Despite a low failure incidence rate of approximately 4%, the developed models demonstrated capacity to discriminate high-risk assets, attaining a peak PR-AUC of 0.153. These results suggest that the framework offers meaningful capability in prioritizing infrequent failure events relevant to practical reliability management. Extending beyond point prediction accuracy, the work illustrated how well-calibrated failure probabilities can be transformed into economic risk metrics to support maintenance optimization. Integration of predictive outputs with an economic objective function enabled identification of an interior optimal inspection threshold (k*), confirming that universal asset inspection is economically suboptimal. When evaluated against alternatives, the inspection portfolio constructed via NPV-based ranking yielded superior economic outcomes relative to both probability-only and random selection strategies. Furthermore, SHAP-based interpretability analysis revealed several engineering-relevant contributors to failure risk, encompassing outage exposure, network density, loading indicators, and environmental stress conditions.

Future research directions for this framework are manifold. First, additional calibration metrics, such as the Brier score and Expected Calibration Error (ECE) could be integrated, alongside cost-sensitive calibration techniques that remain compatible with the maintenance loss functions under consideration. Second, uncertainty in the probability estimates could be explicitly quantified through Bayesian or conformal prediction methods and then incorporated into the selection of robust inspection portfolios. Third, key operational constraints, including crew routing, regional maintenance budgets, and outage coordination, could be embedded within the model, thereby advancing from asset prioritization alone to the generation of fully executable maintenance schedules. Finally, validation on extended temporal datasets or on data drawn from additional utility providers would enable a rigorous evaluation of the framework's generalizability across heterogeneous operational environments.

## Acknowledgements

This research work was funded by the Basic Research Grant (June 2025) from BUET (Bangladesh University of Engineering and Technology) and awarded to Principal Investigator Dr. Md. Saiful Islam. All the computation work of this study was performed in the high-performance server housed at the Institute of Information and Communication Technology (IICT) of BUET in Dhaka, Bangladesh.

## Code Availability

The code used in this study is available from the corresponding author upon reasonable request.

## References

ACER, Annex I to Decision No 23/2020 on the Methodology for Calculating the Value of Lost Load, the Cost of New Entry, and the Reliability Standard, 2020.

Alvarez Quiñones, L.I., Lozano-Moncada, C.A., and Bravo Montenegro, D.A., Machine learning for predictive maintenance scheduling of distribution transformers, *Journal of Quality in Maintenance Engineering*, vol. 29, no. 1, pp. 188–202, 2022.

Bravo M, D.-A., Alvarez Q, L.-I., and Lozano M, C.-A., Dataset of distribution transformers for predictive maintenance, *Data in Brief*, vol. 38, pp. 107454, 2021.

Chen, T. and Guestrin, C., XGBoost: A Scalable Tree Boosting System, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

Friedman, J.H., Greedy function approximation: A gradient boosting machine., *The Annals of Statistics*, vol. 29, no. 5, 2001.

Gorman, W., The quest to quantify the value of lost load: A critical review of the economics of power outages, *The Electricity Journal*, vol. 35, no. 8, pp. 107187, 2022.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K.Q., On Calibration of Modern Neural Networks, 2017.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y., LightGBM: A Highly Efficient Gradient Boosting Decision Tree, *Advances in Neural Information Processing Systems*, vol. 30, pp. 3146–3154, 2017.

Lin, J., Yeo, L.Y., Afrouzi, H.N., Ektesabi, M.M., and Tavalaei, J., From limited data to reliable diagnosis: An interpretable deep learning framework for transformer fault analysis, *International Journal of Electrical Power & Energy Systems*, vol. 172, pp. 111227, 2025.

Lundberg, S.M. and Lee, S.-I., A Unified Approach to Interpreting Model Predictions, *NIPS 2017,* pp. 4768–4777, 2017.

Mahmoud, M.A., Md Nasir, N.R.H., Gurunathan, M., Raj, P., and Mostafa, S.A., The Current State of the Art in Research on Predictive Maintenance in Smart Grid Distribution Network: Fault's Types, Causes, and Prediction Methods—A Systematic Review, *Energies*, vol. 14, no. 16, pp. 5078, 2021.

Sarajčev, P., Jakus, D., and Vasilj, J., Optimal scheduling of power transformers preventive maintenance with Bayesian statistical learning and influence diagrams, *Journal of Cleaner Production*, vol. 258, pp. 120850, 2020.

Vita, V., Fotis, G., Chobanov, V., Pavlatos, C., and Mladenov, V., Predictive Maintenance for Distribution System Operators in Increasing Transformers' Reliability, *Electronics*, vol. 12, no. 6, pp. 1356, 2023.

Wani, S.A., Rana, A.S., Sohail, S., Rahman, O., and Khan, S.A., Advances in DGA based condition monitoring of transformers: A review, *Renewable and Sustainable Energy Reviews*, vol. 149, pp. 111347, 2021.

Zhang, Y., Tang, Y., Liu, Y., and Liang, Z., Fault diagnosis of transformer using artificial intelligence: A review, *Frontiers in Energy Research*, vol. Volume 10-2022, 2022.

## Biographies

**Md. Efatuzzaman Efat** is a student completed his B.Sc. in Mechanical Engineering at the Khulna University of Engineering and Technology (KUET) and currently pursuing a Post Graduate Diploma (PGD) in Business IT at the Institute of Information and Communication Technology (IICT), Bangladesh University of Engineering and Technology (BUET). The challenge of solving real-world problems is what motivates him as a result of a combination of smart computational processes and high-tech engineering systems. Nowadays, his work is dedicated to Material Informatics, the application of AI and ML to the work of mechanical systems, and IoT to the industrial market.

**Irtefa Waseek** holds a Master's degree in Applied Statistics and Data Science and a B.Sc. in Energy Science and Engineering. His academic training integrates statistical modeling, machine learning, and energy systems engineering to address complex infrastructure challenges. He has professional experience as an Assistant Engineer (Energy) and has worked in data-centric roles, including Data Scientist and Data Analyst, across international organizations. His research focuses on Artificial Intelligence, Deep Learning, Smart Grid Systems, Predictive Maintenance, and scalable Big Data Engineering. He is particularly interested in developing robust, data-driven optimization and predictive frameworks to enhance energy reliability, system stability, and intelligent infrastructure automation.

**Sumiya Afrose** is currently pursuing her M.Sc. in Robotics and Mechatronics Engineering at the University of Dhaka. She completed her B.Sc. in Energy Science and Engineering at the Khulna University of Engineering and Technology (KUET). Her research focuses on wearable sensing systems, signal processing, and machine learning. She is particularly motivated by the application of emerging technologies to address real-world challenges across human-technology interaction and intelligent systems design.

**Sayed Md Fazle Rabbi** is presently undertaking his B.Sc. in Electrical and Electronic engineering at Ahsanullah University of Science and Technology (AUST). He has been active in researching interdisciplinary use of emerging technologies, which he started studying in 2021. His areas of research include machine learning, cybersecurity, optical communication, photonic crystal fiber (PCF) and material informatics. The combination of smart computational processes and the high-tech engineering systems to solve the challenges in the real world is what motivates him in particular.

**Md. Saiful Islam** is a Professor at the Institute of Information and Communication Technology (IICT), Bangladesh University of Engineering and Technology (BUET), where he has also served in key academic leadership roles, including Director of IICT. He received his Ph.D. in Computer Science and Engineering from BUET, M.Sc. in CSE from Shanghai University, China, and B.Sc. in EEE from BUET. His research interests include software engineering, databases, optical and wireless communications, cybersecurity, and machine learning, with a strong emphasis on real-world AI applications such as intelligent network intrusion detection, traffic classification, and security analytics for communication systems. He has an active publication record in reputed journals and conferences and supervises graduate research with a strong presence on Google Scholar.