

# Chronological Evaluation of LSTM Forecasting and XGBoost Anomaly Detection in Three-Phase Power Monitoring

Sugeng Widodo<sup>1</sup>, Ardi Pujiyanta<sup>2\*</sup>, Ardiansyah<sup>3</sup>

<sup>1,2,3</sup>Master of Electrical Engineering, Universitas Ahmad Dahlan, Yogyakarta 55164, Indonesia  
[2408057007@webmail.uad.ac.id](mailto:2408057007@webmail.uad.ac.id), [ardipujiyanta@uad.ac.id](mailto:ardipujiyanta@uad.ac.id), [ardiansyah@uad.ac.id](mailto:ardiansyah@uad.ac.id)

## Article Info

### Article history:

Received 2026-08-28

Revised 2026-09-25

Accepted 2026-10-06

### Keyword:

Anomaly Detection,  
LSTM,  
Power Forecasting,  
Three-Phase Monitoring,  
XGBoost.

## ABSTRACT

Three-phase electrical monitoring produces multivariate streams that support both short-term power forecasting and early anomaly detection. This study evaluates a joint LSTM-XGBoost workflow on 30,240 semi-synthetic one-minute records containing 35 attributes and seven anomaly types. The labels were generated by rule-based anomaly injections over a real normal measurement base, and the prediction task is binary anomaly detection. The data were divided chronologically into 14 training days, three validation days, and four test days. An LSTM generated one-step active-power forecasts; its residuals were then combined with raw sensor, first-difference, temporal, and 15-sample history features for XGBoost classification. The LSTM achieved an MAE of 1.4596 W, outperforming persistence (1.5748 W) and the rolling mean (1.9437 W), with Wilcoxon p-values below  $10^{-23}$ . Across ten seeds, the Proposed LSTM-XGB obtained  $F1 = 0.7008 \pm 0.0101$ ,  $ROC-AUC = 0.9530 \pm 0.0074$ , and  $PR-AUC = 0.7913 \pm 0.0094$ . However, Sensor+Difference XGBoost achieved the best threshold-dependent performance, with  $F1 = 0.7163 \pm 0.0081$  and  $PR-AUC = 0.8180 \pm 0.0051$ . These results show that LSTM improves forecasting, whereas residual features provide limited incremental value for anomaly classification beyond direct sensor and first-difference features.



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

## I. INTRODUCTION

Continuous electrical monitoring is increasingly used to expose consumption patterns, power-quality deviations, and unsafe operating conditions. A compact ESP32-based platform has demonstrated that voltage, current, power, and energy can be acquired and delivered to users in real time [1]. Power-quality monitoring designs further show that low-cost IoT devices can support three-phase measurements when sensing and communication are integrated carefully [2].

Three-phase systems require more than aggregate energy reporting because phase imbalance, frequency deviation, voltage sag, and current surges may occur over short intervals[3]. A web-connected three-phase monitoring and control system illustrates the practical value of remote visibility for such parameters [4]. Another ESP32 implementation emphasizes that the monitoring layer must preserve phase-specific measurements rather than collapse them prematurely [5][6]. Company-level power-consumption

monitoring also confirms the operational need for continuous, remotely accessible data [7].

Recent monitoring studies have expanded the communication and presentation layers[8]. A three-phase system based on IoT demonstrates the feasibility of direct online observation [9], while Firebase has been used as a cloud-backed repository for electrical measurements [10]. Node-RED provides another real-time visualization path for power-quality variables [11]. These systems improve observability, but their primary outputs remain dashboards, alarms, or rule-based indicators rather than statistically evaluated predictive models.

Predictive analysis can convert the monitoring stream into forward-looking information. LSTM has been applied to online power prediction because its recurrent state can represent temporal dependence in transformer measurements [12][13]. Physics-informed neural networks offer an interpretable alternative for sensor-based energy-consumption prediction [14]. Conventional machine learning

has also been used to classify electrical loads, demonstrating that engineered features can remain competitive when labels are limited [15].

More distributed learning strategies have appeared in smart plugs and buildings[8]. Semi-supervised federated learning has been proposed for real-time energy optimization while reducing centralized data exchange [16]. Federated learning has likewise been evaluated for environmental monitoring in smart buildings [17]. Predictive energy platforms illustrate that forecasting can also support higher-level decisions beyond local monitoring [18]. Nevertheless, these advances do not by themselves establish whether forecast residuals provide useful evidence for multi-type electrical anomaly detection[19].

Existing three-phase applications often focus on fuzzy control [20], motor-fault monitoring [21], or equipment-specific panels [22]. Their application settings are valuable, but joint evaluation of one-step power forecasting, residual-based anomaly features, time-aware splitting, seed variability, event recall, calibration, and paired significance tests remains limited[23][24]. Consequently, the incremental value of adding a temporal forecaster to a strong sensor-history classifier is frequently unclear.

This study evaluates a joint LSTM-XGBoost workflow rather than assuming that the two models improve both tasks simultaneously. The contributions are: (1) a leakage-aware chronological evaluation on minute-resolution three-phase

measurements; (2) a controlled comparison of persistence, rolling mean, and LSTM forecasting; (3) a feature-level ablation of direct sensor, temporal-history, and LSTM-residual information; and (4) row-level, event-level, calibration, confidence-interval, and paired statistical analyses. The study intentionally reports both positive and negative evidence: LSTM improves active-power forecasting, whereas residual features provide limited incremental value for anomaly detection in this dataset.

## II. METHOD

### A. Dataset and Quality Audit

The experiment used the minute-resolution three-phase monitoring dataset processed by the accompanying notebook. It contains 30,240 rows, 35 columns, and a continuous observation period from 5 June to 25 June 2026. The dataset is a semi-synthetic proof-of-concept derived from 1200 real normal measurements. Anomaly labels were generated as rule-based simulated injections rather than collected as independent field incident labels; therefore, the exact thresholds and event-construction rules must be interpreted together with the dataset-generation procedure. A total of 2,274 rows (7.52%) were anomalous and were grouped into 125 events.

TABLE I  
CHRONOLOGICAL DATA SPLIT

| Split      | Rows   | Days | Anomaly rows | Rate (%) | Events |
|------------|--------|------|--------------|----------|--------|
| Train      | 20,160 | 14   | 1,545        | 7.66     | 81     |
| Validation | 4,320  | 3    | 316          | 7.31     | 18     |
| Test       | 5,760  | 4    | 413          | 7.17     | 26     |

The chronological partition in Table I prevents future observations from entering model fitting. The 14-day training period provides multiple daily operating cycles for parameter estimation, the three-day validation period is used only for checkpoint and threshold selection, and the four-day test period is held out for final evaluation. The split is therefore suitable for a first temporal proof-of-concept, although rolling-origin validation across several temporal windows is needed to establish robustness beyond this single period. The seven anomaly types are current imbalance, current surge,

frequency deviation, power-factor drop, sensor drift, voltage sag, and voltage swell. The classification target is binary (normal versus anomaly); `anomaly_type` is used only for stratified error analysis and is not a seven-class prediction target. In the test period, the anomaly supports are 36, 102, 29, 109, 6, 60, and 71 rows, respectively, grouped into 3, 6, 3, 5, 1, 4, and 4 events. With a one-minute sampling interval, the duration of a contiguous event is calculated from its number of labeled rows and the sampling interval.

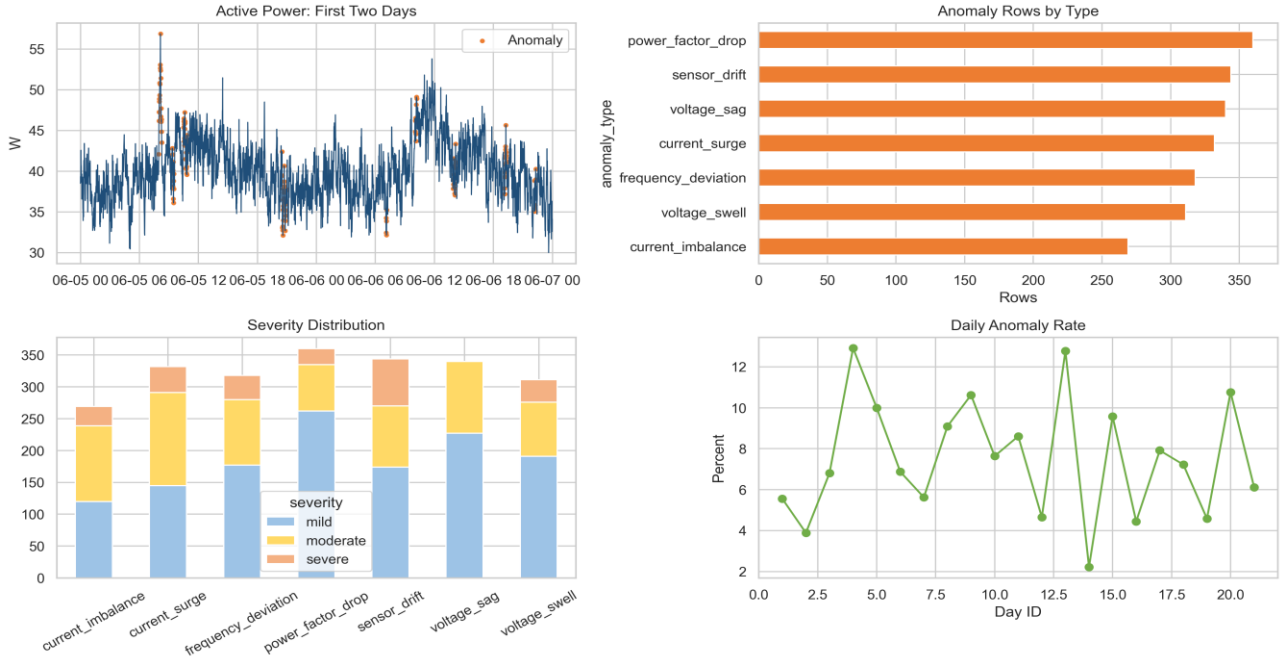


Figure 1. Dataset overview: active-power series, anomaly frequency, severity distribution, and daily anomaly rate.

Figure 1 shows that anomalies occur across the observation period rather than in a single isolated block. Their uneven type and severity distributions motivate chronological evaluation and event-level reporting in addition to ordinary row-level metrics.

### B. Preprocessing and Feature Construction

Raw variables include phase voltages and currents, line voltage, active/reactive/apparent power, power factor, frequency, temperature, humidity, and voltage/current imbalance. First differences are computed from the current value and its immediately preceding observation. For a target time  $t$ , each 15-sample history statistic uses observations ending at  $t-1$ ; the current observation is never included in its own historical mean or standard deviation. Scalers and imputers are fitted on the training period only and then frozen for validation and test. LSTM residuals are generated after the forecaster is trained using only normal training windows, and label, event, severity, split, audit, and provenance columns are excluded from all classifier features.

Let  $X_{(t-L:t-1)}$  denote the ordered multivariate sequence preceding time  $t$ . The sequence-to-one LSTM estimates active power  $y_t$  as

$$\hat{y}_t = f\theta(X_{t-L:t-1}). \quad (1)$$

Forecast-error features are computed only after the forecast is generated. Three residual representations are used:

$$r_t = y_t - \hat{y}_t, \quad a_t = |r_t|, \quad q_t = r_t / (|y_t| + \epsilon). \quad (2)$$

Here, epsilon prevents division by zero and is set to 1.0 W in the implementation. The residual vector  $[r_t, a_t, q_t]$  is concatenated with the selected sensor, first-difference, temporal, and history features before being passed to XGBoost. This design separates forecasting from classification and permits a direct test of whether residual information adds value beyond direct electrical measurements.

### C. LSTM Forecasting Model

The LSTM is a sequence-to-one regressor with a 30-sample look-back window. It contains two recurrent layers with 48 and 24 units, respectively, followed by 0.20 dropout after each recurrent layer, a 16-unit ReLU dense layer, and a one-unit linear output layer. The model is optimized with Adam at an initial learning rate of 0.001, gradient clipping with clipnorm=1.0, and Huber loss on scaled active power. Training uses batch size 128, a maximum of 50 epochs, and shuffle=False. Early stopping monitors validation loss with patience=7, restores the best weights, and uses min\_delta=1e-4. ReduceLROnPlateau uses a factor of 0.5, patience=3, and a minimum learning rate of 1e-5. The recorded run terminates after 26 epochs.

$$MAE = (1/n) \sum_{t=1}^n |y_t - \hat{y}_t|. \quad (3)$$

RMSE and MAPE complement MAE. Metrics are computed for all test observations and separately for normal and anomalous rows. Paired absolute errors are compared using a one-sided Wilcoxon signed-rank test because the sample-wise error distribution need not be Gaussian

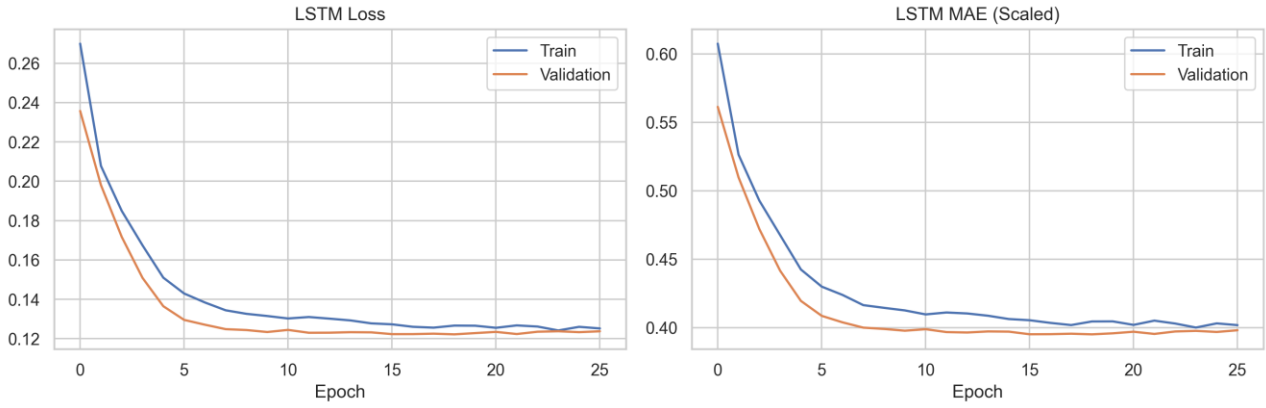


Figure 2. LSTM training and validation loss and scaled MAE across 26 epochs.

The training curves in Figure 2 decrease sharply during the first eight epochs and then stabilize. The small train-validation gap provides no visible evidence of severe overfitting, although the final oscillation supports retaining the best validation checkpoint rather than the last epoch.

#### D. XGBoost Ablation Variants

The anomaly detector is a binary XGBoost classifier. The revised ablation protocol evaluates six variants under the same train/validation/test periods and threshold-selection rule: sensor-only, sensor plus first difference, sensor-history, sensor plus LSTM residual, full LSTM-XGBoost combining sensor-history and residual features, and residual-only as a diagnostic control. Sensor-history features comprise direct electrical variables, first differences, 15-sample history z-scores, and cyclic hour encodings. These variants isolate the incremental contribution of each information source rather than assuming that the full model is automatically superior.

$$c_t = 1 \text{ if } p_t \geq \tau; \text{ otherwise } c_t = 0. \quad (4)$$

The threshold  $\tau$  is selected independently on validation data for each model and seed, then locked before test evaluation. The ten XGBoost seeds (42 to 132 in increments of 10) repeat classifier training, validation threshold selection, and test inference with identical hyperparameters. The LSTM architecture and its trained checkpoint are fixed for the ten-seed XGBoost comparison; hyperparameters are not reselected using test data. Thus, the reported standard deviation measures XGBoost stochastic variability and threshold variation under a fixed temporal protocol.

Algorithm 1. Joint forecasting and anomaly-detection workflow.

1. Input chronologically ordered sensor records and anomaly labels.
2. Audit timestamps, required values, identifiers, and sampling interval
3. Split data chronologically into train, validation, and test periods.

4. Fit preprocessing parameters using the training period only
5. Train the LSTM sequence-to-one active-power forecaster.
6. Generate one-step forecasts and residual-derived features.
7. Build sensor-history, residual-only, and combined XGBoost variants.
8. Select the classification threshold on validation data.
9. Evaluate the locked test set with ten seeds and statistical tests.

#### E. Evaluation and Statistical Analysis

Anomaly detection is evaluated as a binary task with accuracy, precision, recall, specificity, false-positive rate, F1, F2, ROC-AUC, PR-AUC, and Brier score. PR-AUC is interpreted relative to the positive-class prevalence. For operational interpretation, event recall is reported in addition to row-level recall. Detection latency is defined as the time from an event start to the first alarm within that event; the present offline protocol does not include alarm persistence or a streaming alarm state, so latency is treated as a limitation rather than inferred from row-level metrics.

A moving-block bootstrap with block length 30 produces 95% confidence intervals for threshold-dependent classification metrics and paired forecasting effects.

### III. RESULTS AND DISCUSSION

#### A. Power-Forecasting Performance

Table II reports consistent forecasting behavior across all, normal, and anomalous rows. On the full test set, LSTM reduces MAE by 7.31% relative to persistence and by 24.91% relative to the rolling mean. RMSE falls by 7.73% and 25.39%, respectively. The anomaly-only MAE advantage over persistence is smaller (0.57%), indicating that abrupt abnormal transitions remain difficult to predict from preceding context.

TABLE II  
POWER-FORECASTING METRICS

| Scope   | Model        | MAE (W) | RMSE (W) | MAPE (%) |
|---------|--------------|---------|----------|----------|
| All     | Persistence  | 1.5748  | 1.9807   | 3.9663   |
| All     | Rolling mean | 1.9437  | 2.4493   | 4.9040   |
| All     | LSTM         | 1.4596  | 1.8276   | 3.6812   |
| Normal  | Persistence  | 1.5807  | 1.9856   | 3.9781   |
| Normal  | Rolling mean | 1.9420  | 2.4439   | 4.8985   |
| Normal  | LSTM         | 1.4572  | 1.8235   | 3.6710   |
| Anomaly | Persistence  | 1.4994  | 1.9172   | 3.8136   |
| Anomaly | Rolling mean | 1.9662  | 2.5185   | 4.9738   |
| Anomaly | LSTM         | 1.4908  | 1.8796   | 3.8125   |

The paired tests support the aggregate differences. LSTM versus persistence yields  $W=8,201,614$ ,  $p=2.28 \times 10^{-24}$ , and a median absolute-error reduction of 0.0988 W, with a 95% confidence interval of [0.0760, 0.1157] W. LSTM versus the

rolling mean yields  $W=9,763,597$ ,  $p=2.40 \times 10^{-128}$ , with a median absolute-error reduction of 0.3645 W and a 95% confidence interval of [0.3091, 0.4289] W.

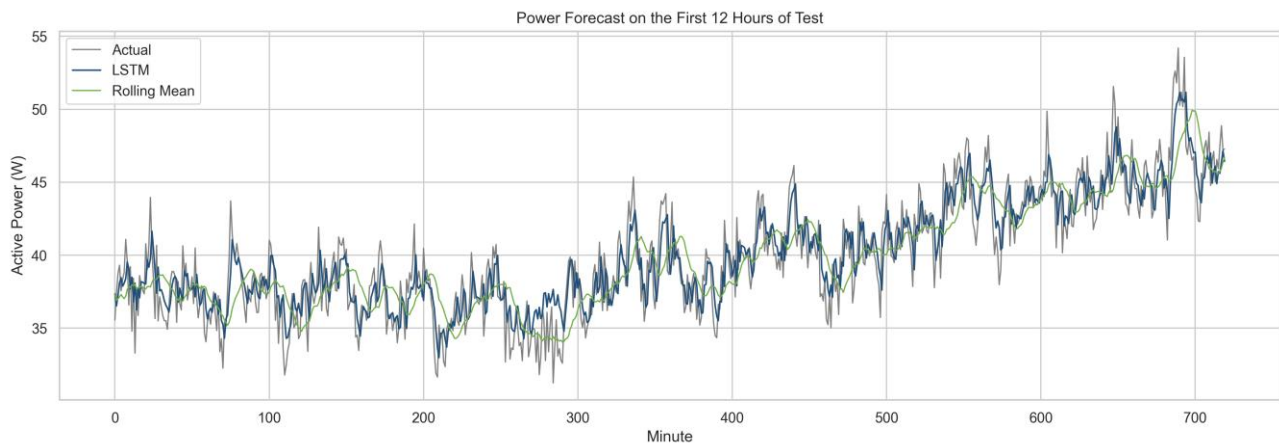


Figure. 3. Actual active power and one-step forecasts during the first 12 hours of the test period

The LSTM forecast remains conservative around the sharpest peaks. This behavior explains why its relative improvement is smaller on anomalous rows, even though it is statistically superior over the complete paired-error distribution.

#### B. Anomaly-Detection Ablation

Table III presents the ten-seed mean and standard deviation. The residual-only classifier has high recall but near-random ROC-AUC and extremely low precision, showing that residual magnitude alone cannot separate abnormal from normal variation. The combined model recovers strong discrimination, but Sensor-History XGB remains marginally better on every mean ranking metric shown. The combined LSTM-XGBoost model does not outperform the Sensor-History XGBoost model. Its mean F1-

score is 0.7011, compared with 0.7088 for Sensor-History XGBoost.

The Proposed LSTM-XGB does not achieve the best threshold-dependent classification performance. Its mean F1-score is  $0.7008 \pm 0.0101$ , whereas Sensor+Difference XGB achieves  $0.7163 \pm 0.0081$  and Sensor-History XGB achieves  $0.6992 \pm 0.0115$ . Although the Proposed LSTM-XGB obtains the highest ROC-AUC among the main sensor-based variants ( $0.9530 \pm 0.0074$ ), the results do not support a claim of overall superiority. Instead, direct sensor and first-difference features provide stronger threshold-dependent anomaly classification in this dataset.

The test set contains 413 anomalous rows among 5,760 evaluated rows, corresponding to an anomaly prevalence of 7.17%. Therefore, the no-skill PR-AUC baseline is approximately 0.0717. The Proposed LSTM-XGB achieves a

mean PR-AUC of  $0.7913 \pm 0.0094$  across ten seeds, demonstrating strong minority-class ranking relative to the prevalence baseline. Nevertheless, PR-AUC does not replace

precision, recall, F1-score, and event-level recall because these metrics describe different operational properties.

TABLE III  
SIX-VARIANT XGBOOST ABLATION RESULTS

| Variant               | F1                                    | F2                                    | Recall                                | Precision                             | ROC-AUC             | PR-AUC                                |
|-----------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------|---------------------------------------|
| LSTM-Residual XGB     | $0.1326 \pm 0.0014$                   | $0.2718 \pm 0.0059$                   | $0.9070 \pm 0.0681$                   | $0.0716 \pm 0.0004$                   | $0.5092 \pm 0.0037$ | $0.0785 \pm 0.0009$                   |
| Proposed LSTM-XGB     | $0.7008 \pm 0.0101$                   | $0.6689 \pm 0.0097$                   | $0.6494 \pm 0.0150$                   | $0.7626 \pm 0.0345$                   | $0.9530 \pm 0.0074$ | $0.7913 \pm 0.0094$                   |
| Sensor+Difference XGB | <b><math>0.7163 \pm 0.0081</math></b> | <b><math>0.6936 \pm 0.0183</math></b> | <b><math>0.6797 \pm 0.0294</math></b> | $0.7601 \pm 0.0371$                   | $0.9471 \pm 0.0023$ | <b><math>0.8180 \pm 0.0051</math></b> |
| Sensor+Residual XGB   | $0.6787 \pm 0.0108$                   | $0.6324 \pm 0.0103$                   | $0.6051 \pm 0.0144$                   | $0.7743 \pm 0.0380$                   | $0.9007 \pm 0.0023$ | $0.7583 \pm 0.0033$                   |
| Sensor-History XGB    | $0.6992 \pm 0.0115$                   | $0.6686 \pm 0.0225$                   | $0.6501 \pm 0.0328$                   | $0.7602 \pm 0.0434$                   | $0.9481 \pm 0.0046$ | $0.7896 \pm 0.0103$                   |
| Sensor-Only XGB       | $0.6864 \pm 0.0139$                   | $0.6372 \pm 0.0085$                   | $0.6082 \pm 0.0089$                   | <b><math>0.7885 \pm 0.0358</math></b> | $0.9022 \pm 0.0020$ | $0.7657 \pm 0.0080$                   |

Table III presents the mean and standard deviation across ten XGBoost seeds for six classification variants. Sensor+Difference XGB achieves the highest F1-score ( $0.7163 \pm 0.0081$ ), F2-score ( $0.6936 \pm 0.0183$ ), recall ( $0.6797 \pm 0.0294$ ), and PR-AUC ( $0.8180 \pm 0.0051$ ). The Proposed LSTM-XGB obtains the highest ROC-AUC among the main sensor-based variants ( $0.9530 \pm 0.0074$ ), but its F1-score is lower than Sensor+Difference XGB. The residual-only classifier performs poorly, with an F1-score of 0.1326 and a PR-AUC of 0.0785. These results indicate that LSTM

residuals provide strong ranking information when combined with temporal sensor features, but they do not improve the threshold-dependent classification performance beyond the simpler sensor-difference representation. These results indicate that LSTM residuals improve ranking performance when combined with temporal sensor features, but they do not improve threshold-dependent classification beyond direct sensor and first-difference features.

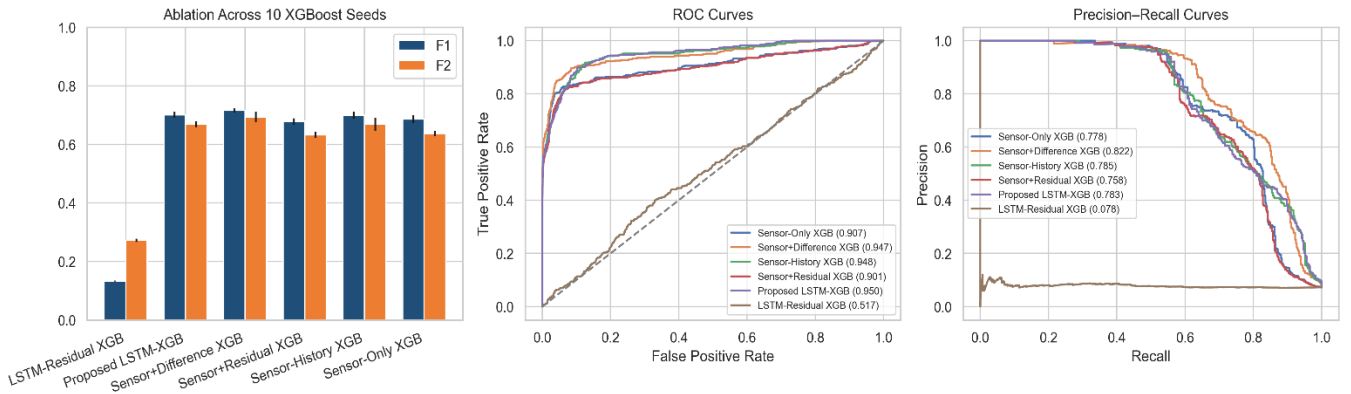


Figure 4. Six-variant XGBoost ablation results with F1/F2 comparison and ROC/precision-recall curves.

Figure 4 shows that the residual-only classifier is clearly separated from the other variants. Sensor+Difference XGB provides the strongest precision-recall performance, whereas the Proposed LSTM-XGB achieves competitive ROC-AUC. This difference confirms that ranking performance and threshold-dependent classification performance do not necessarily lead to the same model selection.

### C. Selected-Run Reliability

For the representative seed-42 run, the selected threshold is 0.520. The confusion matrix contains 5,212 true negatives, 105 false positives, 147 false negatives, and 266 true positives. The false-positive rate is calculated as  $105 /$

$(5,212 + 105) = 1.98\%$ , while precision and recall are 0.7170 and 0.6441, respectively. The resulting F1-score is 0.6786. These values provide a threshold-dependent operational view that complements ROC-AUC and PR-AUC. The corresponding confusion matrix is shown in Figure. 5

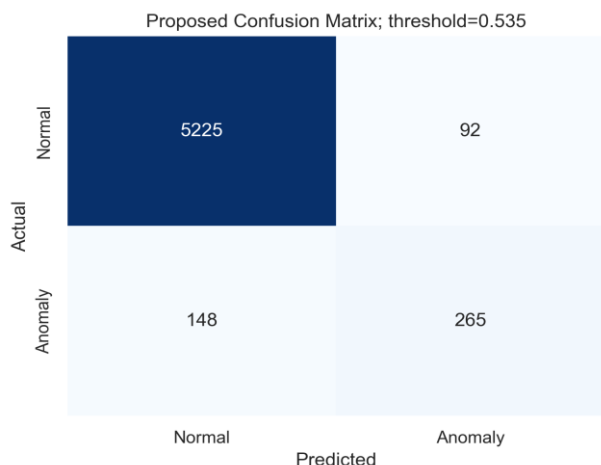


Figure 5. Confusion matrix of the Proposed LSTM-XGB model at threshold 0.520.

TABLE IV  
MOVING-BLOCK BOOTSTRAP 95% CONFIDENCE INTERVALS

| Metric    | Estimate | 2.5%  | 97.5% |
|-----------|----------|-------|-------|
| Precision | .7423    | .5772 | .8932 |
| Recall    | .6416    | .4840 | .7821 |
| F1        | .6883    | .5511 | .7892 |
| F2        | .6595    | .5117 | .7766 |
| ROC-AUC   | .9497    | .9154 | .9761 |
| PR-AUC    | .7740    | .6406 | .8691 |

TABLE V  
RECALL BY ANOMALY TYPE ON THE TEST PERIOD

| Anomaly type        | Rows | Total duration (min) | Events | TP | FN | Row recall | Detected events | Missed events | Event recall |
|---------------------|------|----------------------|--------|----|----|------------|-----------------|---------------|--------------|
| Current imbalance   | 36   | 36                   | 3      | 36 | 0  | 1.000      | 3               | 0             | 1.000        |
| Current surge       | 102  | 102                  | 6      | 31 | 71 | 0.304      | 6               | 0             | 1.000        |
| Frequency deviation | 29   | 29                   | 3      | 25 | 4  | 0.862      | 3               | 0             | 1.000        |
| Power-factor drop   | 109  | 109                  | 5      | 97 | 12 | 0.890      | 5               | 0             | 1.000        |
| Sensor drift        | 6    | 6                    | 1      | 0  | 6  | 0.000      | 0               | 1             | 0.000        |
| Voltage sag         | 60   | 60                   | 4      | 58 | 2  | 0.967      | 4               | 0             | 1.000        |
| Voltage swell       | 71   | 71                   | 4      | 19 | 52 | 0.268      | 3               | 1             | 0.750        |

Table V shows that detection performance varies substantially across anomaly types. Current imbalance, frequency deviation, power-factor drop, and voltage sag achieve complete event-level recall, although some individual anomalous rows remain undetected. Current surge is detected in all six events despite a relatively low row recall of 0.304, indicating that the model captures at least one anomalous row

Table IV shows that discrimination remains high after block-resampling, but the uncertainty of threshold-dependent minority metrics is materially wider. The lower F1 bound of 0.5511 cautions against relying only on a single confusion matrix. In contrast, the ROC-AUC interval remains above 0.91.

#### D. Event-Level Recall and Failure Modes

Row recall does not fully represent operational usefulness because an alarm can be useful even if it captures only part of an event. Because the prediction target is binary, the seven anomaly types are evaluated using one-versus-rest post-hoc stratification. For each type, Table V reports the number of anomalous rows, total duration, number of events, true positives, false negatives, row recall, detected events, missed events, and event recall. Total duration is calculated using the 60-second sampling interval. The model detects 24 of 26 test events (92.31%) and 266 of 413 anomalous rows (64.41%). Per-type false positives are not assigned because the same normal-row false positive may contribute to more than one one-versus-rest view. The overall binary test set contains 105 false positives. Table V summarizes the row-level and event-level detection results for each anomaly type during the test period.

in every surge event. Voltage swell is detected in three of four events, with a row recall of 0.268 and an event recall of 0.750. Sensor drift is the clearest failure mode because all six anomalous rows in its only test event are missed. This result indicates that event-level recall can be substantially higher than row-level recall and should therefore be reported alongside row-based metrics.

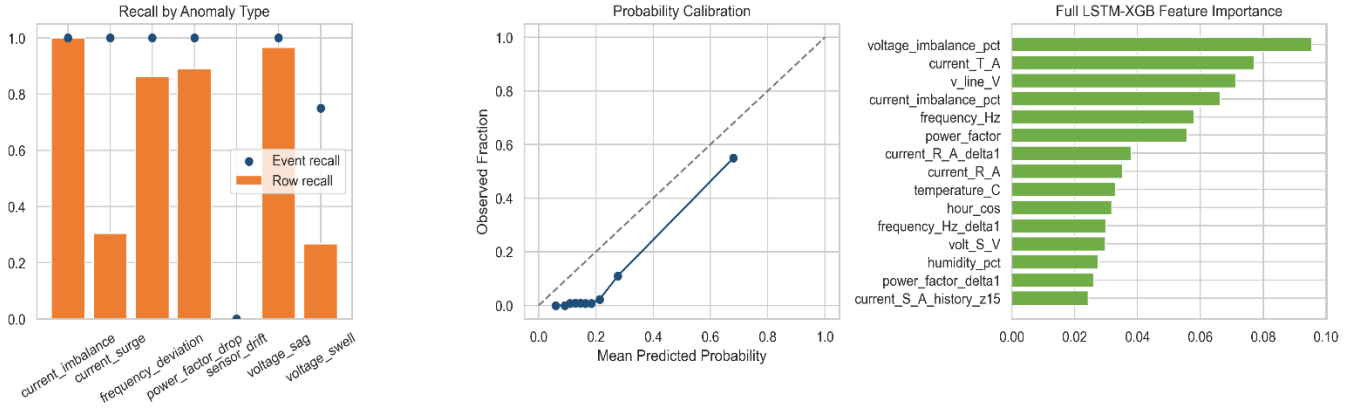


Figure 6. Anomaly-type recall, probability calibration, and leading feature importance.

TABLE VI  
ONE-VERSUS-REST ANOMALY-TYPE CLASSIFICATION RESULTS

| Anomaly type        | Precision | Recall | F1-score |
|---------------------|-----------|--------|----------|
| Current imbalance   | 0.0970    | 1.0000 | 0.1769   |
| Current surge       | 0.0836    | 0.3039 | 0.1311   |
| Frequency deviation | 0.0674    | 0.8621 | 0.1250   |
| Power-factor drop   | 0.2615    | 0.8899 | 0.4042   |
| Sensor drift        | 0.0000    | 0.0000 | 0.0000   |
| Voltage sag         | 0.1563    | 0.9667 | 0.2691   |
| Voltage swell       | 0.0512    | 0.2676 | 0.0860   |

Table VI reports the one-versus-rest precision, recall, and F1-score for each anomaly type during the test period. In this analysis, one anomaly type is treated as the positive class, while all remaining observations are treated as the negative class. The results show substantial variation across anomaly types. Power-factor drop achieves the highest F1-score (0.4042), supported by a precision of 0.2615 and a recall of 0.8899. Voltage sag obtains the highest recall among the difficult anomaly types (0.9667), although its precision remains relatively low at 0.1563. Current imbalance achieves perfect recall (1.0000) but low precision (0.0970), indicating that many predicted alarms do not correspond to this specific anomaly type. Sensor drift is not detected, resulting in zero precision, recall, and F1-score. Current surge and voltage swell also show low F1-scores because of their relatively low recall and the large number of false positives.

Because the same normal-row false positives may contribute to more than one one-versus-rest evaluation, type-specific precision and F1-score should be interpreted cautiously. These results are used for post-hoc anomaly-type analysis and do not represent a seven-class classification model.

#### E. Calibration and Feature Importance

As shown in Fig. 6, anomaly-type recall varies substantially across categories, while the calibration curve and feature-importance plot provide additional diagnostic evidence. The selected model obtains a Brier score of 0.05055. The calibration bins lie mostly below the ideal

diagonal, meaning predicted anomaly probabilities tend to exceed the observed fraction within each bin. Thresholded discrimination is useful, but the raw probabilities should therefore be recalibrated before being interpreted as operational risk.

Feature importance explains the ablation result. Voltage imbalance (0.0907), phase-T current (0.0752), current imbalance (0.0698), line voltage (0.0677), frequency (0.0572), and power factor (0.0553) dominate. In contrast, relative, absolute, and signed LSTM residual features have importances of only 0.0038, 0.0032, and 0.0018. The combined model therefore behaves primarily as a sensor-history classifier.

#### F. Comparison with Related Monitoring Approaches

The result complements rather than replaces existing monitoring architectures. Generator-panel monitoring demonstrates reliable three-phase acquisition and remote presentation [25], while sensor-specific three-phase designs show how current and voltage channels can be instrumented [26]. Multifunction meters based on low-cost controllers likewise provide a practical acquisition layer [27]. The present study adds a time-aware analytical layer and quantifies uncertainty, but it does not evaluate hardware reliability or communication latency.

Equipment-level studies report that IoT panels can support pump monitoring [22] and motor protection [28]. Induction-motor monitoring also illustrates the value of continuous phase measurements for fault awareness [29]. A panel-focused implementation demonstrates how electrical variables can be exposed to operators [30]. These applications are plausible deployment contexts for the Combined workflow, yet field validation is still required before transferring the current threshold and probability calibration.

#### G. Limitations and Practical Implications

Five limitations bound the conclusions. First, the evaluation covers one 21-day dataset and should not be generalized to different sensors, loads, or seasonal conditions without external validation. Second, the number of events is

small for sensor drift and several other types. Third, the output does not include embedded-device latency, memory, packet loss, or long-term sensor degradation. Fourth, the current study uses one fixed chronological split; rolling-origin temporal splits should be added in future work to test stability across changing operating periods. Fifth, detection latency was not estimated because the offline experiment does not model alarm persistence, suppression, or streaming state. XGBoost importance is associative and may distribute credit among correlated electrical features.

For practical use, the sensor-history model is currently the simpler anomaly detector because it slightly improves accuracy and avoids dependence on the LSTM pipeline. The LSTM remains valuable when a power forecast is itself an operational output. A joint deployment is therefore justified when both forecasting and anomaly detection are required, not because residual features alone improve classification. Future work should evaluate probability calibration, anomaly-type-specific thresholds, longer field recordings, and models that learn residual dynamics rather than appending three scalar residual features.

#### IV. CONCLUSION

The results show that LSTM improves one-step active-power forecasting, achieving an MAE of 1.4596 W and an RMSE of 1.8276 W. For anomaly detection, Sensor+Difference XGB achieves the best threshold-dependent performance, with an F1-score of 0.7163 and a PR-AUC of 0.8180. The Proposed LSTM-XGB provides strong ranking performance, with a ROC-AUC of 0.9530, but its residual features do not improve classification performance beyond the simpler sensor-difference representation. Therefore, the contribution of the proposed workflow is not universal superiority, but an empirical evaluation showing that LSTM is useful for forecasting while direct temporal sensor transformations remain more effective for anomaly classification in this dataset.

#### ACKNOWLEDGMENTS

This research received a research grant from the Ministry of Research, Technology, and Higher Education (Ristekdikti) of the Republic of Indonesia, under contract number NOMOR: 056/PTM/LPPM.UAD/IV/2026.

#### REFERENCES

- [1] H. J. El-Khozondar *et al.*, "A smart energy monitoring system using ESP32 microcontroller," *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, vol. 9, no. March, p. 100666, 2024, doi: 10.1016/j.prime.2024.100666.
- [2] H. Prasetyo, Y. Sugiarto, A. B. Raharja, A. O. Rachmawan, B. D. Mahadi, and F. P. Pratama, "Design of IoT Enabled Three Phase Power Quality Monitoring Unit Based on EmonLibCM Library," *KONSTELASI: Konvergensi Teknologi dan Sistem Informasi*, vol. 4, no. 2, pp. 1–14, 2024, doi: 10.24002/konstelasi.v4i2.10231.
- [3] E. Viciano, F. M. Arrabal-Campos, A. Alcayde, R. Baños, and F. G. Montoya, "All-in-one three-phase smart meter and power quality analyzer with extended IoT capabilities," *Measurement*, vol. 206, p. 112309, 2023, doi: https://doi.org/10.1016/j.measurement.2022.112309.
- [4] C. D. Hantoro and S. Setiawidayat, "Monitoring and Control of 3 Phase Electrical Energy Internet of Things (IoT) Based," *European Journal of Electrical Engineering and Computer Science*, vol. 7, no. 3, pp. 80–86, 2023, doi: 10.24018/ejece.2023.7.3.522.
- [5] W. Yuniarto, I. I. S. S. R. Man, M. Diponegoro, and E. E., "Rancang Bangun Sistem Monitoring Dan Kontrol Energi Listrik Pada Beban 3 Fasa Menggunakan Esp32 Berbasis Internet of Think (Iot)," *Jurnal Poli-Teknologi*, vol. 22, no. 1, pp. 30–38, 2023, doi: 10.32722/pt.v22i1.5102.
- [6] B. Mallala, R. R. Manyam, J. Saini, M. F. N. Tajuddin, and S. B. Thanikanti, "IoT enabled smart power quality analysis in three phase electrical systems with practical implementation," *Scientific Reports*, vol. 15, no. 1, p. 39665, 2025, doi: 10.1038/s41598-025-23317-z.
- [7] R. Kango, S. Suhaedi, and F. A. Hasanuddin, "Implementation of The Internet of Things for Monitoring The Company's Electrical Power Consumption," *Journal of Asian Multicultural Research for Economy and Management Study*, vol. 2, no. 1, pp. 16–22, 2021, doi: 10.47616/jamrems.v2i1.72.
- [8] Y. Li, D. Qin, H. V. Poor, and Y. Wang, "Introducing edge intelligence to smart meters via federated split learning," *Nature Communications*, vol. 15, no. 1, p. 9044, 2024, doi: 10.1038/s41467-024-53352-9.
- [9] N. Hidayati and S. Nengsih, "Implementation of the Internet of Things in the Design and Development of a Three-Phase Power Monitoring System," *Circuit: Jurnal Ilmiah Pendidikan Teknik Elektro*, vol. 9, no. 2, pp. 4–8, 2025, doi: 10.22373/crc.v9i2.29309.
- [10] M. Z. Z. Muhtadi, S. Ranolat, and Pujiyanto, "Sistem Monitoring Energi Listrik 3 Fase Berbasis IoT dan Firebase," *Elposys: Jurnal Sistem Kelistrikan*, vol. 12, no. 2, pp. 132–137, 2025, doi: 10.33795/elposys.v12i2.7206.
- [11] D. Fadhillah Rahman Riyanto, D. Widjajanto, and Isdawimah, "Sistem Monitoring Kualitas Daya Listrik Real-Time Berbasis IoT dengan Node-RED," *Prosiding Seminar Nasional Teknik Elektro*, vol. 11, pp. 1–7, 2025.
- [12] D. A. Pratama and S. Nabila, "Application of Deep Learning LSTM in Online Power Prediction on Three-Phase Power Transformer," *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 3, 2022, doi: 10.47065/bits.v4i3.2690.
- [13] I. Ullah, S. Muhammad Hasanat, K. Aurangzeb, M. Alhussein, M. Rizwan, and M. S. Anwar, "Multi-horizon short-term load forecasting using hybrid of LSTM and modified split convolution," *PeerJ Computer Science*, vol. 9, p. e1487, Sep. 2023, doi: 10.7717/peerj-cs.1487.
- [14] S. Hossain, A. Rahman, S. Ahmed, S. Enam, and M. T. Ahmed, "Interpretable Physics-Informed Neural Networks for energy consumption prediction using IoT sensors," *Array*, vol. 28, no. October, p. 100552, 2025, doi: 10.1016/j.array.2025.100552.
- [15] S. Salma, F. Dewanta, and M. Abdillah, "Klasifikasi Beban Listrik Dengan Machine Learning Menggunakan Metode K-Nearest Neighbor," *RESISTOR (Elektronika Kendali Telekomunikasi Tenaga Listrik Komputer)*, vol. 5, no. 2, p. 163, 2022, doi: 10.24853/resistor.5.2.163-172.
- [16] M. M. Hossen, A. Nabil Sayed, A. Nhlabatsi, M. S. Islam, and F. Bensaali, "Smart plug for real-time energy monitoring and optimization through micro-moments with semi-supervised federated learning," *Results in Engineering*, vol. 29, no. February 2025, p. 108568, 2026, doi: 10.1016/j.rineng.2025.108568.
- [17] P. Ventura, M. Khodamoradi, R. Costa, P. Figueiras, and R. Jardim-Gonçalves, "Real-Time Environmental Monitoring in Smart Buildings Using Federated Learning," *Procedia Computer Science*, vol. 263, pp. 680–687, 2025, doi: 10.1016/j.procs.2025.07.081.
- [18] V. K. Mololoth, C. Åhlund, and S. Saguna, "EnergyFlow: Predictive trading platform for decentralized energy exchange," *Sustainable Energy, Grids and Networks*, vol. 45, no. November 2025, p. 102074, 2026, doi: 10.1016/j.segan.2025.102074.
- [19] A. Akagic and I. Džafić, "Enhancing smart grid resilience with deep learning anomaly detection prior to state estimation," *Engineering*

- Applications of Artificial Intelligence*, vol. 127, p. 107368, 2024, doi: <https://doi.org/10.1016/j.engappai.2023.107368>.
- [20] R. S. Siregar and M. Ikhsan, "Implementasi Internet of Thing (IOT) untuk Monitoring dan Pengaman Beban Listrik 3 Phase Menggunakan Fuzzy Sugeno," *Jurnal Algoritma*, vol. 22, no. 1, pp. 1013–1025, 2025, doi: 10.33364/algoritma/v.22-1.2343.
- [21] E. P. Wicaksono and S. I. Haryudo, "Prototipe Sistem Monitoring Gangguan Motor Tiga Fasa Berbasis Internet of Things Prototype of a Three-Phase Motor Fault Monitoring System Based on IoT," *Telekontran: Jurnal Ilmiah Telekomunikasi, Kendali dan Elektronika Terapan*, vol. 11, no. 2, pp. 105–114, 2023.
- [22] M. Ihsan, A. R. Al Muttaqi, U. A. Ahmad, J. S. Wicaksana, and F. Azmi, "Development of 3 Phase Pump Panel Monitoring System based on Internet of Things (Case Study PERUMDAM)," *[CEPAT] Journal of Computer Engineering: Progress, Application and Technology*, vol. 1, no. 01, p. 23, 2022, doi: 10.25124/cepat.v1i01.4846.
- [23] M. F. Guato Burgos, J. Morato, and F. P. Vizcaino Imacaña, "A Review of Smart Grid Anomaly Detection Approaches Pertaining to Artificial Intelligence," *Applied Sciences*, vol. 14, no. 3, 2024, doi: 10.3390/app14031194.
- [24] Y. Zheng, F. Chen, H. Yang, and S. Su, "Edge Computing Based Electricity-Theft Detection of Low-Voltage Users," *Frontiers in Energy Research*, vol. Volume 10-2022, 2022, doi: 10.3389/fenrg.2022.892541.
- [25] G. Babys, A. Mashar, and W. Mursanto, "Rancang Bangun Sistem Pemantauan Daya Listrik Generator Sinkron Tiga Fasa Berbasis Internet of Things," *Jurnal Teknik Energi*, vol. 11, no. 2, pp. 37–42, 2023, doi: 10.35313/energi.v11i2.4682.
- [26] Fasha Nur Fauzan, Sri Utami, and Mulki Rezka Budi Pratama, "Monitoring Sistem Kelistrikan Tiga Fasa Berbasis IoT dengan Sensor ACS712 dan ZMPT101B," *Jurnal Teknik Energi*, vol. 13, no. 1, pp. 24–29, 2024, doi: 10.35313/v13i1.5246.
- [27] Erry Yadie, Marson Ady Putra, and Muhammad Zaini Aqmal, "Multifunction Meter 3 Fase Menggunakan Sensor Pzem-004t Berbasis Nodemcu Esp8266," *PoliGrid*, vol. 6, no. 1, pp. 41–49, 2025, doi: 10.46964/poligrid.v6i1.57.
- [28] D. K. Putra, F. Faizah, and A. Musadek, "Rancang Bangun Sistem Proteksi adn Kontrol Kecepatan Motor 3 Phase Berbasis Internet Of Thing (IoT)," *PROSIDING Seminar Nasional Inovasi Teknologi Penerbangan (SNITP)*, vol. 6, no. 1, pp. 1–13, 2022.
- [29] R. Febri, A. Basuki, W. Suluh, and A. Pambudi, "Monitoring System Motor Induksi 3 Fasa Berbasis Internet of Things," *Journal of Electrical Electronic Control and Automotive Engineering (JEECAE)* 47 *JEECAE*, vol. 5, no. 2, pp. 47–50, 2020.
- [30] A. D. Dharmawan, L. Subiyanto, and A. T. Nugraha, "Elektriese: Jurnal Sains dan Teknologi Elektro Implementasi Sistem Monitoring pada Panel Listrik," *Jurnal Sains dan Teknologi Elektro Implementasi Sistem Monitoring pada Panel Listrik*, vol. 12, no. 02, pp. 82–91, 2023.