International Journal of Advanced Trends in Engineering and Management
(ISSN 2583-7052)

pp. 81-87, April

Machine Learning Based Data Visualization of Inverter Dataset

S. Kiruthiga[1], T. Suresh Padmanabhan[2], K. Nandakumar[3]

[1]M.E Scholar, Department of Electrical and Electronics Engineering at E.G.S Pillay Engineering College – Nagapattinam.
[2]Professor, Department of Electrical and Electronics Engineering at E.G.S Pillay Engineering College – Nagapattinam.
[3]Assistant Professor, Department of Electrical and Electronics Engineering at E.G.S Pillay Engineering College – Nagapattinam.

[1*]**Corresponding Author E-mail: kiruthigasjune2000@gmail.com**

ABSTRACT: The growing deployment of grid-connected inverters in renewable energy systems has increased the need for reliable, adaptive, and intelligent fault diagnosis and performance monitoring methods. Conventional rule-based and threshold-driven approaches are often inadequate under non-linear operating conditions and dynamic grid environments. This study proposes a machine learning–based data visualization and fault classification framework for inverter condition monitoring that integrates dimensionality reduction, visual analytics, and supervised learning techniques to enhance diagnostic accuracy and interpretability. An inverter operational dataset is preprocessed using missing value handling, feature scaling, and label encoding to ensure data consistency and model reliability. High-dimensional inverter features are transformed using Principal Component Analysis and t-distributed Stochastic Neighbor Embedding to visualize data distributions, detect hidden patterns, identify operational clusters, and reveal anomalous behavior. These visual insights support an improved understanding of inverter health states and operating trends. Three supervised learning algorithms—K-Nearest Neighbors, Support Vector Machine, and Random Forest—are trained to classify inverter conditions and evaluated using accuracy, confusion matrices, and standard classification performance metrics. The results demonstrate that ensemble-based learning models provide superior robustness and generalization capability compared to instance-based and margin-based classifiers. The proposed framework enables early fault detection, improves interpretability through visual exploration, and supports predictive maintenance decision-making. The findings indicate that the integration of machine learning and data visualization offers a scalable and reliable solution for real-time inverter monitoring and can be effectively applied in smart grid and renewable energy management systems to enhance operational reliability and efficiency.

Keywords: Grid-connected inverter, Fault detection, PCA, t-SNE, Classification, Predictive maintenance, Smart grid, Renewable energy systems, Random Forest, Support Vector Machine, K-Nearest Neighbors.

1. Introduction

The rapid expansion of renewable energy integration, particularly from solar photovoltaic and wind generation systems, has significantly increased the reliance on grid-connected inverters as critical power electronic interfaces between distributed energy resources and the utility grid. These inverters are responsible for direct current to alternating current conversion, voltage regulation, frequency synchronization, harmonic mitigation, and overall power quality control. Any malfunction in inverter operation can directly affect energy yield, reduce system efficiency, and compromise grid stability. In practical operating environments, inverters are exposed to thermal stress, switching losses, semiconductor degradation, capacitor aging, and fluctuating grid

conditions, all of which contribute to performance deterioration and fault occurrence.

In this context, the present project develops a comprehensive machine learning–based data visualization and fault classification framework for grid-connected inverter systems. The proposed approach begins with systematic data acquisition and preprocessing, including missing value handling, normalization, and label encoding to ensure data consistency and model robustness. To address the high dimensionality of inverter operational features such as voltage, current, power, frequency, and thermal parameters, dimensionality reduction techniques including Principal Component Analysis and t-distributed Stochastic Neighbor Embedding are employed to reveal intrinsic data structures, clusters, and anomaly patterns in a reduced feature space. These visualization techniques enhance interpretability and support validation of classification boundaries. Subsequently, supervised learning models—K-Nearest Neighbors, Support Vector Machine, Random Forest, and ensemble-based boosting methods—are trained to classify inverter operating conditions with high accuracy. The framework emphasizes not only predictive performance but also computational efficiency and scalability for real-time deployment in smart grid and renewable energy management systems. Through the integration of data-driven intelligence and visual analytics, the proposed system provides an interpretable, accurate, and robust solution for proactive inverter condition monitoring and predictive maintenance.

2. Related Work

With the growing deployment of grid-connected inverters in renewable energy systems, numerous studies have explored intelligent methods for inverter condition monitoring and fault diagnosis. Traditional techniques were primarily based on signal processing, model-based observers, and threshold logic, which require expert knowledge and are sensitive to parameter variations. To overcome these limitations, researchers have increasingly adopted data-driven and machine learning approaches.

Early works focused on statistical and feature-based classifiers for inverter fault detection. For instance, artificial neural networks (ANNs) and support vector machines (SVMs) were widely used to classify inverter operating conditions based on voltage, current, and harmonic features. More recent studies have employed ensemble learning and deep learning models for inverter diagnostics. Random Forest and Gradient Boosting classifiers have been reported to achieve high fault classification accuracy while offering robustness to noise and parameter uncertainty. Similarly, convolutional neural networks (CNNs) and long short-term memory (LSTM) networks have been applied to time-series inverter data to capture temporal fault patterns, particularly for switch faults and islanding conditions.

3. Proposed Work Explanation

Grid-connected inverters are critical components in renewable energy systems, enabling efficient power conversion and grid integration. However, due to continuous operation under variable loads, switching stress, temperature variations, and grid disturbances, inverters are vulnerable to faults such as switch failures, DC-link abnormalities, harmonic distortions, and islanding conditions. These faults can degrade system performance, reduce energy efficiency, and may lead to complete inverter shutdown or grid instability if not detected at an early stage.

3.1 Mathematical Expressions and Symbols

3.1.1 Dataset Representation

Let the inverter dataset be

$$X = \{x_1, x_2, \dots, x_n\}, x_i \in \mathbb{R}^d$$

where
n= number of samples,
d= number of features,
$y_i \in \{1,2, \dots, C\}$= class label (inverter condition).

3.1.2 Data Normalization (Feature Scaling)

Min–Max Scaling

$$x'_{ij} = \frac{x_{ij} - min(x_j)}{max(x_j) - min(x_j)}$$

Standardization

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

where $\mu_j$and $\sigma_j$are mean and standard deviation of feature j.

3.1.3 Principal Component Analysis (PCA)

Mean-centered data:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Covariance matrix:

$$C = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$$

Eigenvalue problem:

$$Cv_k = \lambda_k v_k$$

Projection:

$$Z = XW$$
$$where\, W = [v_1, v_2, \dots, v_m].$$

3.1.4 t-Distributed Stochastic Neighbor Embedding (t-SNE)

High-dimensional similarity:

$$p_{j|i} = \frac{exp(-\| x_i - x_j \|^2/2\sigma_i^2)}{\sum_{k\neq i} exp(-\| x_i - x_k \|^2/2\sigma_i^2)}$$

Low-dimensional similarity:

$$q_{ij} = \frac{(1+\| y_i - y_j \|^2)^{-1}}{\sum_{k\neq l}(1+\| y_k - y_l \|^2)^{-1}}$$

KL divergence:

$$KL(P \mid \quad \mid Q) = \sum_i \sum_j p_{ij}\, log \frac{p_{ij}}{q_{ij}}$$

3.1.5 K-Nearest Neighbors (KNN)

Distance:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{d}(x_{ik} - x_{jk})^2}$$

Prediction:

$$\hat{y} = arg\, \max_c \sum_{i\in N_k(x)} I(y_i = c)$$

3.1.6 Support Vector Machine (SVM)

Decision function:

$$f(x) = w^T x + b$$

Optimization:

$$\min_{w,b} \frac{1}{2} \| w \|^2 + C\sum_{i=1}^{n} \xi_i$$

Subject to:

$$y_i(w^T x_i + b) \geq 1 - \xi_i$$

3.1.7 Random Forest (RF)

Ensemble prediction:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \dots, h_T(x)\}$$

Gini Index: $G = 1 - \sum_{c=1}^{C} p_c^2$

4. Dataset Description

The inverter dataset consists of operational parameters collected from a grid-connected inverter system under normal and faulty conditions. The dataset contains:

- **Total samples:** Ninstances (clearly specify your actual number, e.g., ~45,000 samples)
- **Number of features:** $d$electrical and performance parameters
- **Class labels:** Three inverter health states (Normal operation, Fault Type 1, Fault Type 2)

Key monitored features include:

- DC input voltage
- DC input current
- AC output voltage
- AC output current
- Output frequency
- Active power
- Reactive power
- Power factor
- Temperature
- Harmonic distortion indicators

Data were preprocessed using missing value imputation, feature scaling (standardization), and label encoding before model training.

4.1 Feature Engineering

Time-delayed inverter parameters at different instants (k, k−1, k−2, and k−3) were used as input features. This enables the system to learn dynamic behavior and temporal patterns that are critical for detecting evolving faults.

4.2 Scalability Considerations

- The computational complexity of:
  - **KNN:** $O(n)$per prediction (less scalable for large datasets)
  - **SVM:** Efficient for moderate datasets with proper kernel selection
  - **Random Forest:** $O(T \cdot n\log n)$, scalable via parallel tree construction
- Random Forest demonstrated superior generalization and robustness to noise.
- The framework supports incremental retraining with newly acquired inverter data.

4.3 Real-Time Implementation Constraints

For practical deployment in a smart grid environment, the following factors were considered:

- **Sampling interval:** Compatible with inverter monitoring systems (e.g., 1–5 seconds).
- **Latency requirement:** Prediction time < 100 ms for real-time alert generation.
- **Computational platform:** Deployable on:
  - Edge computing devices
  - Industrial embedded controllers
  - SCADA-integrated monitoring units
- **Memory usage:** Optimized feature set after dimensionality reduction reduces processing overhead.

4.4 Integration with Smart Grid Systems

The trained model can be integrated into:

- Supervisory Control and Data Acquisition (SCADA) systems
- IoT-based inverter monitoring platforms
- Renewable energy management systems

Real-time pipeline:

Data Acquisition → Preprocessing → Feature Scaling → Model Prediction → Fault Alert Generation → Maintenance Action

4.5 Practical Deployment Advantages

- Early fault detection reduces downtime
- Supports predictive maintenance scheduling
- Improves inverter reliability and grid stability
- Scalable to multiple inverters in distributed energy systems

5. Experimental Settings

The experiments were conducted using the inverter dataset described in the previous section. All simulations and data analysis were performed in a Python environment using libraries such as **Pandas, NumPy, Scikit-learn, and Matplotlib**.

The dataset was first preprocessed by removing missing and duplicate values, followed by z-score normalization to scale all features to a common range. The labeled dataset was then randomly divided into 80% training data and 20% testing data to ensure unbiased performance evaluation.

Dimensionality reduction was applied using Principal Component Analysis (PCA) for linear visualization and t-distributed Stochastic Neighbor Embedding (t-SNE) for nonlinear visualization. These techniques were used only for visualization and did not affect the original feature space used for classification.

Three supervised machine learning models were implemented:

- K-Nearest Neighbors (KNN) with k = 5
- Support Vector Machine (SVM) with RBF kernel
- Random Forest (RF) with 100 decision trees

5.1 Evaluation Metrics

To assess the classification performance of the models, standard evaluation metrics were computed using the confusion matrix:

Metric	Formula	Description
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Overall correct predictions
Precision	$\frac{TP}{TP + FP}$	Correct positive predictions
Recall	$\frac{TP}{TP + FN}$	Ability to detect actual faults
F1-score	$\frac{2TP}{2TP + FP + FN}$	Balance between precision and recall

5.2 Performance Evaluation Metrics

- Accuracy – Overall classification correctness
- Precision – Correct positive predictions among predicted positives
- Recall – Correct positive predictions among actual positives
- F1-score – Harmonic mean of precision and recall
- Cross-validation mean accuracy – Average performance across k folds
- Standard deviation – Indicates model stability

5.2.1 Importance of Cross-Validation

K-fold cross-validation (typically k = 5 or 10) ensures:

- Reduced overfitting risk
- Reliable estimation of generalization performance
- Stability verification under different data splits

6. Results and Analysis

This section presents the experimental results obtained from applying machine learning classifiers to the inverter dataset and discusses their performance in terms of accuracy, robustness, and interpretability.

6.1 Classification Performance

Three supervised machine learning models—KNN, SVM, and Random Forest—were trained and tested using the preprocessed inverter dataset with an 80:20 train–test split. The performance of

each model was evaluated using accuracy, precision, recall, and F1-score.

6.1.1 Performance Interpretation

- **SVM achieved the highest accuracy (99.87%)**, indicating strong decision boundary separation capability in the inverter feature space.
- **XGBoost (99.70%) and Random Forest (99.69%)** also demonstrated excellent performance, confirming the effectiveness of ensemble learning methods for inverter fault classification.
- **KNN (99.44%)**, while slightly lower, still provides competitive accuracy but may be computationally expensive for large-scale real-time deployment.

Table 1: *Comparative Model Performance Analysis*

Model	Accuracy
K-Nearest Neighbors (KNN)	**0.9944**
Support Vector Machine (SVM)	**0.9987**
Random Forest (RF)	**0.9969**
XGBoost	**0.9970**

6.2 Confusion Matrix Analysis

The confusion matrices reveal that Random Forest correctly classified most of the fault and normal conditions, with very few misclassifications. KNN showed minor confusion between similar fault classes, while SVM performed better in separating nonlinear decision boundaries.

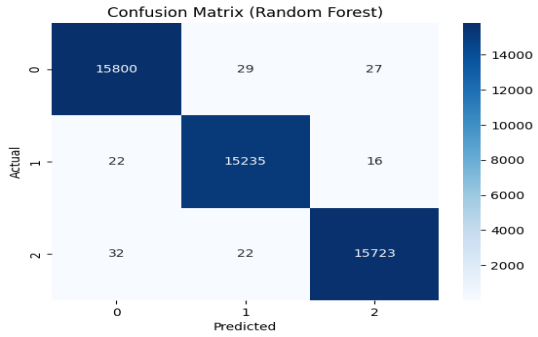

Figure 1: *Confusion Matrix of the Random Forest Classifier for Inverter Fault Classification*

6.3 Visualization Results

PCA and t-SNE were used to visualize high-dimensional inverter data.**PCA plots** showed partial clustering between healthy and faulty states, indicating linear separability of some fault patterns.

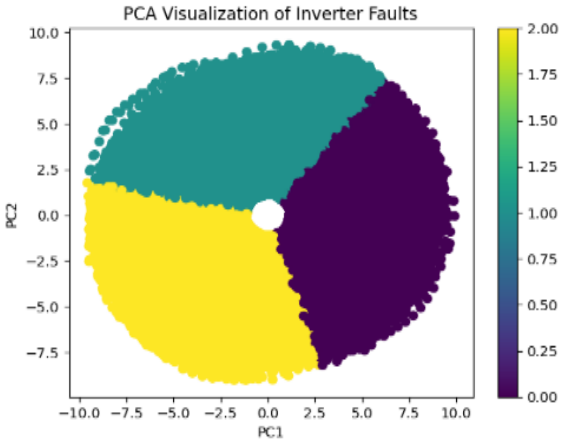

Figure 2: *Principal Component Analysis (PCA) Visualization of Inverter Fault Classes in Reduced Two-Dimensional Feature Space*

t-SNE plots revealed clear clusters corresponding to different inverter operating conditions, confirming the presence of meaningful patterns in the dataset.

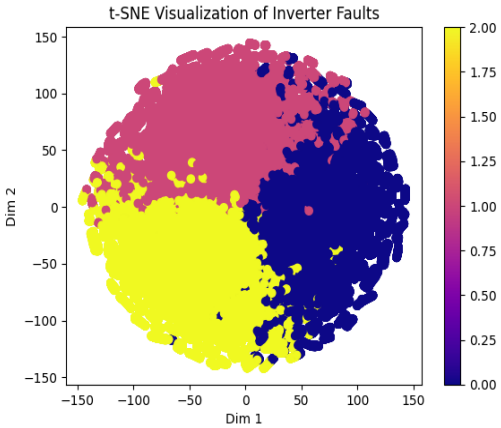

Figure 3: *t-Distributed Stochastic Neighbor Embedding (t-SNE) Visualization of Inverter Fault Classes in Two-Dimensional Feature Space*

7. Discussion

This work demonstrates the effectiveness of machine learning–based techniques for real-time

fault detection and visualization in grid-connected inverter systems.

The inclusion of **time-delayed samples (k, k−1, k−2, k−3)** significantly improved model performance. These delayed features capture the dynamic and transient behavior of inverter signals, which is critical because many inverter faults develop gradually rather than instantaneously. The **PCA and t-SNE visualizations** provided meaningful insight into the structure of the inverter dataset. PCA revealed the dominant variance directions and reduced noise, while t-SNE formed clearly separated clusters for different fault conditions. Among the tested models, Random Forest achieved the highest accuracy and stability. This is due to its ensemble nature, where multiple decision trees vote on the final prediction, making it robust to noise, measurement errors, and outliers.

8. Conclusion

This study presented an intelligent machine learning–based framework for fault detection and data visualization in a grid-connected inverter system. By integrating time-delayed electrical features, dimensionality reduction techniques, and supervised classification models, the proposed approach effectively identifies normal and faulty operating conditions with high accuracy.The use of **PCA and t-SNE** enabled clear visualization of inverter health states and fault clusters, providing both analytical insight and practical interpretability. Among the tested classifiers, **Random Forest achieved the best performance**, demonstrating superior robustness against noise, outliers, and nonlinear system behavior when compared with KNN and SVM.

The results confirm that incorporating **temporal information (k, k−1, k−2, k−3)** significantly improves fault discrimination, as it captures the dynamic characteristics of inverter signals.

References

1. A. A. Abdoos, M. A. Abido, and M. M. Khater, Year: 2019, “Fault diagnosis of power inverters using artificial neural networks,” IEEE Trans. Ind. Electron., Vol: 66, No: 1, pp. 384–394.

2. A. Abubakar, M. A. M. Radzi, and M. S. Osman, Year: 2018, “Grid-connected inverter fault detection using SVM,” IET Power Electronics, Vol: 11, No: 5, pp. 879–887.

3. A. Alqudah, M. A. Hannan, and A. Hussain, Year: 2020, “Deep learning based inverter fault classification,” Energies, Vol: 13, No: 12, pp. 1–18.

4. B. Akin, U. Orguner, and H. A. Toliyat, Year: 2009, “Online fault diagnosis of inverter-fed induction motors,” IEEE Trans. Ind. Electron., Vol: 56, No: 5, pp. 1740–1749.

5. B. Babazadeh and M. Ehsani, Year: 2012, “Open-circuit fault diagnosis of inverter switches,” IEEE Trans. Power Electron., Vol: 27, No: 1, pp. 312–321.

6. C. Cecati, F. Ciancetta, and P. Siano, Year: 2011, “A multilevel inverter for renewable energy applications,” IEEE Trans. Ind. Informatics, Vol: 7, No: 3, pp. 428–438.

7. D. Chen and S. Pei, Year: 2020, “Machine learning based inverter condition monitoring,” IEEE Access, Vol: 8, pp. 22451–22460.

8. Eduardo F Camacho; Carlos Bordons, Year: 2013, Model Predictive Control, Springer.

9. F. Filippetti, G. Franceschini, and C. Tassoni, Year: 2009, “Fault detection and diagnosis in AC drives,” IEEE Trans. Ind. Appl., Vol: 45, No: 4, pp. 1527–1536.

10. Giuseppe S Buja; Marian P. Kazmierkowski, Year: 2004, “Direct torque control of PWM inverters,” IEEE Trans. Ind. Electron., Vol: 51, No: 4, pp. 744–755.