A Study on the Impact of Categorical Alarm Data in Power Estimation and Anomaly Detection of Photovoltaic Inverters

Ruiz Amantegui Jorge [1], Hai-Canh Vu [2], Phuc Do[3]

[1] *Université de Lorraine, CNRS, CRAN, F-54000 Nancy, France*
jorge.ruiz-amantegui@univ-lorraine.fr,
[2] *Université de Technologie de Compiègne, CS 60319, CEDEX, 60203 Compiègne, France*
hai-canh.vu@utc.fr
[3] *SyCoIA, IMT Mines Ales, Ales, France*
phuc.do@mines-ales.fr

ABSTRACT

This study investigates the impact of incorporating categorical inverter data into power-forecasting and anomaly-detection frameworks. Three forecasting models are evaluated on their ability to estimate power output on a large dataset coming from a fleet of multiple photovoltaic plants, over one hundred inverters and an approximate total of 33.2 installed MW. The forecasting models employed are Multi-Layer Perceptron, Long Short-Term Memory, and Extreme Gradient Boosting. Two encoding strategies for categorical alarm codes are compared: one-hot encoding and entity embeddings. Anomaly detection is performed by analysing residuals between predicted and measured power output. By systematically evaluating the integration of categorical inverter data into PV monitoring models, this work addresses an important gap in the literature and provides a foundation for future research exploring advanced methods for exploiting categorical operational data in photovoltaic systems.

1. INTRODUCTION

The (International Energy Agency, 2025) (IEA) has set a net-zero emissions target by 2050, which requires rapid expansion of clean energy technologies, along with improvements in energy efficiency and electrification. Among renewable technologies, solar photovoltaics (PV) exhibit the fastest growth across all projected scenarios .

To address reliability challenges in PV systems, a wide range of prognostics and health management (PHM) techniques have been developed. PHM approaches in PV systems span from physics-based models to advanced data-driven techniques. For instance, (Sheppard et al., 2024) investigate physics-based models for fault detection in the DC collector field of PV systems.

Beyond analytical approaches, data-driven methods have gained significant attention. Classical machine learning (ML) techniques such as Artificial Neural Networks (ANN) (Onal, 2022), Support Vector Machines (SVM) (Hashemi, Taheri, Cretu, & Pouresmaeil, 2021), and Random Forests (RF) (Yao, Kang, Zhou, Abusorrah, & Al-Turki, 2021; Liebermann, Um, Hwang, & Schlüter, 2021) have demonstrated strong performance in fault detection and diagnosis tasks.

More recently, deep learning (DL) architectures have further expanded PHM capabilities. Work by (Chang & Han, 2024) provides a comprehensive review of DL applications in PV system PHM, covering a wide range of architectures such as deep neural networks (DNN), deep autoencoders (DAE), convolutional neural networks (CNN), generative adversarial networks (GAN), and more.

Despite the rapid development of PHM techniques for PV systems, anomaly detection remains challenging. A major issue is the scarcity of labelled failure data. By definition, PV systems operate under normal conditions for the vast majority of their lifetimes, resulting in highly imbalanced datasets in which abnormal events represent only a small fraction of observations.Table 1 shows in more detail the imbalanced nature of the dataset, after labelling the data into normal and abnormal periods, less than 3% of the data is abnormal. Because of this imbalance, a common strategy in anomaly detection is to model normal behaviour and identify deviations from this learned baseline as potential faults.

Table 1. Class distribution of the dataset.

Class	Count	Percentage (%)
Normal	20 093 811	97.53
Shutdown	508 426	2.47

Beyond data imbalance, another under explored aspect concerns the use of categorical information generated directly by PV inverters. To the best of the authors' knowledge, existing PHM approaches rely on continuous SCADA measurements, while the systematic integration of inverter alarm codes as predictive features remains unexplored. Modern inverters, such as Huawei devices, continuously transmit numerical codes describing their operational state and potential fault conditions.

These numeric codes provide detailed information about the inverter's state, grid conditions, input voltages and currents, temperatures, and other operational parameters. They represent a potentially valuable yet insufficiently exploited source of diagnostic information.

In this work, we investigate the impact of integrating categorical inverter data into power-forecasting and anomaly-detection models. Specifically, we compare three forecasting architectures for predicting inverter active power output: Multi-Layer Perceptron (MLP), Extreme Gradient Boosting (XGB), and Long Short-Term Memory (LSTM).

In addition, we evaluate two embedding strategies for incorporating categorical alarm data: one-hot encoding and entity embeddings.

In the context of this work, the alarm messages provided by the inverters are represented only by numeric codes and do not include any associated textual description. Due to this dataset limitation, the semantic meaning of the alarms is unavailable during model development. Consequently, all alarm codes are treated as categorical identifiers without prior interpretation. Any potential relationships between them must therefore be inferred directly from the data. Techniques such as entity embeddings allow the model to learn latent relationships between these categorical variables based on their statistical behaviour in the dataset.

To detect anomalies, the deviation between the predicted and measured power output is analysed to generate O& M warnings.

The analysis is conducted using real operational data collected from several PV systems through SCADA monitoring platforms. Continuous features used for model training are obtained from field-deployed sensors, while categorical data is transmitted directly by the inverters.

In addition, O& M records documenting system failures are available for labelling purposes. These textual records are processed using regular expressions and keyword matching to extract structured failure information.

The remainder of this document is organized as follows. Section 2 goes over the related work and our scientific contribution. Section 3 describes the dataset, including the number of PV plants and inverters, the total operational time, and the features considered. Following, section 4 introduces the embedding techniques for encoding categorical data and the forecasting algorithm.

Section 5 details the proposed methodology, data preprocessing, and the labelling procedure. It also describes the systematic model training process for power prediction, including hyperparameter tuning, as well as the residual analysis used to generate user warnings.

Section 6 presents the experimental results and compares the models using forecasting and anomaly detection metrics. Finally, Section 7 summarizes the main conclusions of the study.

2. RELATED WORK AND SCIENTIFIC CONTRIBUTION

2.1. Anomaly detection in PV systems using AI

Two common approaches to anomaly detection in PV systems rely either on image-based inspection of PV modules or on ML methods that use operational time-series data collected from SCADA systems. While imaging techniques can effectively identify physical defects at the module level, they typically require specialised inspection equipment and are not suited for continuous monitoring. Therefore, this review focuses on data-driven methods based on operational measurements.

Several studies have proposed ML models that learn the normal electrical behaviour of PV systems and detect anomalies from deviations between predicted and measured signals. For example, (Ibrahim, Alsheikh, Awaysheh, & Alshehri, 2022) compared AutoEncoder-LSTM, Facebook Prophet, and Isolation Forest models trained on inverter operational measurements. Their study, however, was conducted on a relatively small dataset consisting of only two inverters and thirteen known anomalies. Similarly, (De Benedetti, Leonardi, Messina, Santoro, & Vasilakos, 2018) trained an ANN to estimate the expected AC power output of a PV system from irradiance and temperature measurements, and identified anomalies through residual analysis. Their method was validated on a PV plant containing 13 inverters.

Other works combine predictive modelling with additional anomaly detection techniques. In his work, (Marangis et al., 2024) propose a predictive maintenance workflow that combines performance modelling with XGBoost, anomaly detection with a one-class SVM, and trend forecasting with the Prophet algorithm, and validated it on a dataset of four inverters and 37 labelled faulty days. Similarly, (Syamsuddin, Adhi, Kusumawardhani, Prahasto, & Widodo, 2024) developed an LSTM autoencoder framework trained on SCADA measurements to detect anomalies through reconstruction errors using operational data collected from 18 inverters.

2.2. Research Gap

Despite the growing body of research on anomaly detection and predictive maintenance in PV systems, several limitations remain in the current literature.

First, many existing studies are evaluated on relatively small datasets. For example, (Ibrahim et al., 2022) evaluate their approach on a dataset composed of only two inverters and a small number of labelled anomalies, while other works rely on single-plant deployments with a limited number of monitored devices (Marangis et al., 2024; Syamsuddin et al., 2024). While these studies demonstrate the feasibility of ML approaches for anomaly detection in PV systems, the scalability and robustness of these methods remain insufficiently explored in large-scale PV fleets with hundreds of inverters.

Second, most anomaly detection methods in PV systems rely almost exclusively on continuous operational variables. However, modern inverter monitoring systems continuously generate categorical information such as alarms and warning codes. These alarms contain potentially valuable diagnostic information related to abnormal operating conditions. Despite their potential relevance for PHM, such categorical signals have received little attention in the anomaly detection literature for PV systems.

These limitations highlight the need for studies that evaluate anomaly detection methods on large-scale, real-world datasets and explore the integration of categorical operational signals generated by PV inverters.

The use of categorical or event-based operational data for predictive maintenance has already demonstrated value in other industrial domains. For instance, (Bezerra et al., 2019) highlight that alarm and event logs constitute an important source of operational information that is often underutilized in industrial systems.

Beyond exploratory analysis, several studies have incorporated categorical operational records directly into PHM frameworks. In their work, (Luo, Li, Zhang, Zhao, & Lim, 2008), propose a framework for constructing degradation models that predict equipment failures using categorical data generated by manufacturing systems, thereby extracting degradation patterns from discrete operational records. Similarly, (Gutschi, Furian, Suschnigg, Neubacher, & Voessner, 2019) investigate the use of machine log messages, event logs, and operational information to estimate the probability of machine breakdown in manufacturing systems. Their work integrates data mining, feature extraction, and ML methods, demonstrating that machine failures can indeed be predicted using such event-based information.

Despite these advances in other industrial domains, the exploitation of categorical alarm data for PHM remains largely unexplored in PV systems. In particular, inverter-generated alarm streams have rarely been considered as predictive features in anomaly detection or performance modelling tasks. Addressing this gap requires evaluating anomaly detection methods on large-scale PV datasets while investigating how categorical alarm information can be integrated into data-driven monitoring frameworks.

2.3. Scientific Contribution

To address the aforementioned limitations, this work proposes an anomaly detection framework for PV inverter monitoring that integrates categorical alarm information with operational time-series data and evaluates its performance on a large-scale dataset. The main scientific contributions of this work are the following:

1. Systematic exploration of categorical inverter alarm data for PV monitoring. This study investigates an underutilized source of information generated by inverter monitoring systems: categorical alarm signals including error codes, warning codes, and inverter state messages. While these signals are routinely recorded in operational databases, they have rarely been incorporated into data-driven predictive maintenance approaches in the PV domain. In this work, an exploratory analysis of these categorical variables is conducted, and their integration into ML-based monitoring models is evaluated, thereby establishing a baseline for future research on the use of categorical operational data in PV systems.
2. Evaluation on a large-scale PV dataset. The proposed methodology is validated using operational data from 126 PV inverters deployed across multiple plants, representing several megawatts of installed capacity. The scale of this dataset exceeds that of many previously published studies in PV anomaly detection, enabling a more robust evaluation of model performance under realistic operating conditions.
3. Systematic comparison of ML models and categorical feature encoding strategies. This study compares the performance of several ML and DL architectures for power prediction and anomaly detection. In addition, two different encoding strategies for categorical alarm data are compared: one-hot encoding and embedding-based representations, providing insights into the impact of categorical data representation on model performance.

The categorical data considered in this work consist of numerical identifiers corresponding to inverter warning codes, error codes, and operational state messages. Although textual descriptions for these codes were not available, these signals are expected to reflect internal inverter conditions and operational events.

This study addresses these limitations by evaluating anomaly detection methods on a large-scale real-world PV dataset while

systematically investigating the use of categorical inverter alarm signals as predictive features for data-driven monitoring models.

The proposed framework can be visualized in Figure 1. Where dark blue boxes denote data objects, intermediate artifacts, or outputs; yellow boxes represent processes; light blue boxes denote optional preprocessing operations, of which at most one may be applied. E.E. stands for Entity Embeddings, O.H. stands for One Hot encoding, e is the prediction residual and DUL is the daily upper limit threshold for said residual used in the warning logic.

3. DATASET

The dataset used in this study consists of real-world operational measurements collected at the inverter level from PV plants located in Germany. A total 13 PV plants and 126 Conergy inverters are analysed. On average, each inverter provides approximately 8.3 years of operational data, resulting in an aggregated observation period of roughly 1045 inverter-years. The operational lifespan of the inverters varies considerably, ranging from a minimum of 164 days to a maximum of 3492 days. A summary of the inverter ages and operational periods across the different plants is presented in Table 3.

For each inverter, the continuous features collected are: irradiance, inverter temperature, module temperature, internal voltage, and degradation over time.

Furthermore, grid-connected solar inverters can transmit information about their current state as a numeric code. These codes have been recorded in our dataset into five different categories, which can be seen next to their cardinality in Table 2 .

Table 2. Available categorical features and their cardinality

Category	Cardinality
Error code 1	68
Error code 2	211
Warning code 1	11
Warning code 2	13
State of the inverter	21

Preliminary analysis of inverter monitoring data supports its potential value. The distribution of alarm codes across the two classes (normal and abnormal) reveals that certain error codes appear more frequently during abnormal conditions, as can be seen in Figure 2.

These observations suggest that categorical inverter alarm data may provide valuable complementary information for PHM applications in PV systems. Importantly, this information is generated internally by the inverter monitoring system and therefore does not require additional sensors or hardware. As a result, leveraging this data can improve monitoring capabilities without increasing instrumentation costs, thereby enhancing its practical value for industrial deployments.

Table 3. Summary of PV plants, operational lifetime, and installed power

Plant ID	Inverters	Lifetime (yr)	Power (MW)
1	1	9.6	0.28
16	62	8.5	17.36
19	3	8.8	0.84
27	8	8.7	2.24
31	5	8.9	1.40
37	9	8.4	2.14
38	5	8.9	1.40
39	8	8.9	1.44
40	3	7.0	0.84
41	6	8.9	1.68
42	8	8.9	1.80
43	4	8.9	0.88
57	4	8.8	0.90

4. EMBEDDING AND FORECASTING METHODS EMPLOYED

4.1. Categorical Encoding Methods

Categorical variables cannot be directly processed by most ML models and must therefore be converted into numerical representations. Two encoding strategies are evaluated in this work.

One-hot encoding converts a categorical variable with k possible values into k binary features, where only one element is equal to 1, and the others are 0. This representation prevents the introduction of artificial ordinal relationships between categories and is widely used in classical ML pipelines. However, it produces high-dimensional and sparse feature spaces when categorical variables have many unique values.

Entity embeddings (Guo & Berkhahn, 2016), provide a learned dense representation of categorical variables. Instead of representing each category as a sparse binary vector, categories are mapped to low-dimensional continuous vectors that are learned during neural network training . This approach allows the model to capture latent similarities among categories and typically scales better with high-cardinality categorical variables.

4.2. Forecasting Models

Three forecasting models are evaluated for predicting inverter power output.

Multi-Layer Perceptron (MLP) is a feedforward neural network trained using backpropagation (Rumelhart, Hinton, & Williams, 1986) . MLPs are effective at modelling non-linear relationships between input variables, but do not explicitly capture temporal dependencies in time-series data.

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network designed to model sequential data

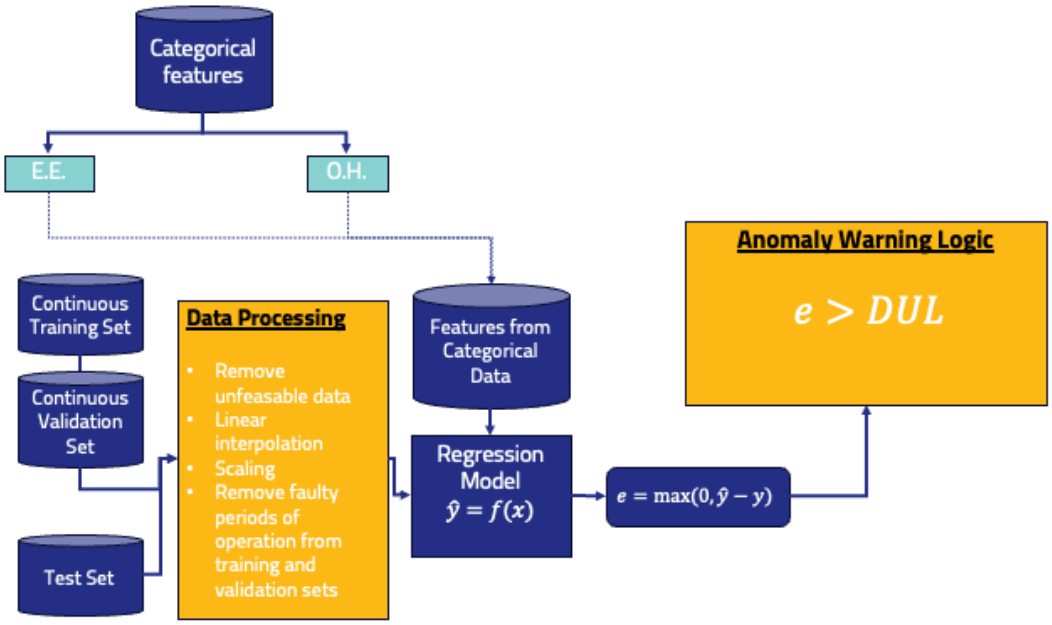

Figure 1. Anomaly detection flowchart

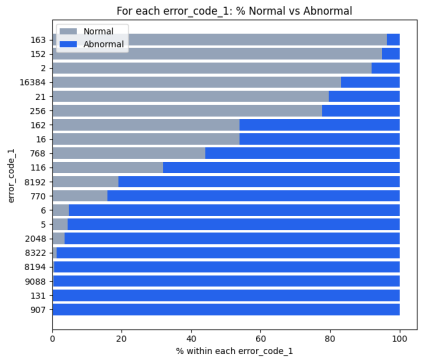

Figure 2. Comparative distribution of inverter error codes during normal and abnormal operation. The heterogeneity and concentration of specific alarms during abnormal periods suggest that categorical alarm information is valuable for predicting abnormal behaviour.

(Hochreiter & Schmidhuber, 1997). Their gated architecture allows the model to capture temporal dependencies while mitigating the vanishing gradient problem present in traditional RNNs.

XGBoost is a gradient-boosted decision tree algorithm that builds an ensemble of trees sequentially, where each tree corrects the residual errors of the previous ones (Chen & Guestrin, 2016). Due to its strong performance on tabular data and efficient implementation, it is widely used as a baseline in many ML applications.

5. METHODOLOGY

5.1. Pre-processing and exploratory data analysis (EDA)

Data pre-processing is an essential step in any ML project. This statement holds especially true when dealing with real data, as it often contains noise: outliers, periods of missing data, and unreasonable data that defies the physical systems under study. These disturbances often arise from errors in the system's installed sensors, and it is of utmost importance to address them before using this data to train ML or DL models.

Our dataset showed gaps and periods of infeasible data, such as large irradiance values ($\approx 1000 W m^{-2}$) at nighttime for more than one consecutive month.

Finally, the nighttime data points are removed from the dataset, as they constitute a significant portion of the data and do not provide useful information for model training.

To deal with missing data, small gaps are filled with a linear interpolation. Any gap of 3 or more data points (30 minutes) is left untouched. Then, any days with missing values are removed from the dataset.

The continuous features are scaled between 0 and 1 using equation 1. Scaling is important for models like MLPs or LSTMs to prevent a single feature from dominating the training process.

$$x_{\text{scaled}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

The max and min values for x are taken from the maximum and minimum values of each feature for each inverter model. The inverter fleet consists of inverters of different sizes from the same brand. These inverters operate at different voltages and currents; therefore, they are grouped by model before scaling the continuous features.

5.2. Labelling

The labels used in this study were derived from three complementary sources:

- Information from the O& M ticketing system, which records interventions and reported failures in the PV system.
- A rule-based detection method designed to identify complete inverter shutdowns.
- Manual inspection of power and irradiance time series.

The O&M records consist of the maintenance logs with the following categories: affected component, start timestamp, end timestamp, and description. To efficiently read through and process the maintenance tickets, they are filtered using search keywords.

If either the affected component or the initial timestamps are missing from the record, the ticket is discarded. Using these records, we define five different categories of tickets: inverter failures, isolation defects, system failures, planned maintenance actions and communication failures.

The first three cases are labelled as failures, and the last two are excluded from the analysis.

The second method of labelling consists of a rule-based approach designed to identify inverter shutdowns. Consecutive periods of operation during which the inverter output power is equal to zero while irradiance measurements exceed $5Wm^{-2}$ are detected. When such periods persist for longer than 12 hours, they are labelled as failures. This duration threshold was chosen to avoid short temporary drops in production caused by sensing errors, communication issues, or even curtailment.

The final labelling step consisted of manual validation of candidate anomalous events. After training and evaluating each model configuration, the days flagged as anomalous were visually inspected using the corresponding power production, irradiance, and inverter voltage time series. A day was labelled as a failure when clear abnormal operating patterns were observed. These included sustained underperformance relative to the available irradiance, complete inverter shutdown during periods of sufficient irradiance, and characteristic patterns associated with string disconnections or derating.

This validation procedure was repeated for each model architecture and each input configuration tested in this study (continuous variables only, one-hot encoded categorical variables, and entity embeddings). By applying the same inspection process to the anomalies detected by each model configuration, the labelling procedure ensured that no particular model or feature representation was favoured in constructing the ground-truth dataset.

5.3. Training methodology

A regression model is trained independently for each inverter. For each device, the first year of operation is used as the training dataset to represent normal operating conditions. The following two months of data are used as the validation set.

Both datasets are filtered to resemble normal operation. First, data points where either the power output or the irradiance is below 0.05 are removed, as illustrated in Figure 3. This step removes measurements corresponding to night-time conditions or very low irradiance levels, where the relationship between power and irradiance is not informative.

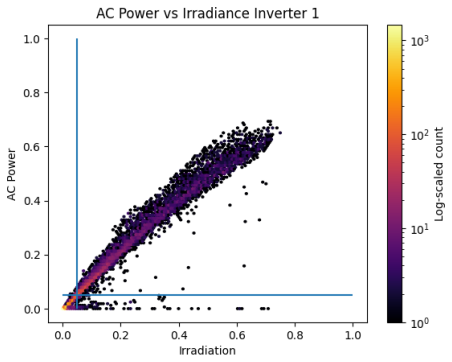

Figure 3. Removed data points to resemble normal operation

Next, any periods labelled as failures are removed from both the training and validation datasets. This preprocessing step ensures that the model is trained only on data representing normal inverter behaviour. Such an approach, where anomalous or failure events are excluded from the training data, is commonly adopted in predictive maintenance studies for PV systems (De Benedetti et al., 2018).

The model hyperparameters are optimized using the Optuna

library to minimize MAE on the validation set when estimating inverter output power. Once the optimization process is complete, the configuration achieving the lowest MAE on the validation set is selected and used to generate predictions on the test set. This procedure is repeated independently for each inverter. The hyperparameters optimized depend on the model architecture being trained. The overall training procedure is illustrated in Figure 4. During hyperparameter tuning, validation MAE values lower than 0.020 were commonly achieved. These values are substantially lower than those observed on the test set (see Table 4 and Figure 6), as expected, since the hyperparameters are explicitly optimized on the validation data.

5.4. Anomaly warning logic

Anomaly detection is performed by comparing the actual power in the test dataset to the power predicted by the model. If the difference between the predicted and measured power is sufficiently large, an anomalous behaviour alarm is raised.

Finding these thresholds and generating the alarms is a process divided into three steps:

- Clipping the residuals to zero
- Finding the threshold of anomalous behaviour by the inverter
- Raising daily alarms

The first step is to compute the residuals as the difference between the expected power and actual power at each instance t, as seen in 2:

$$r_t = \max\big(0, \hat{y}_t - y_t\big) \tag{2}$$

where r_t is the residual, $\hat{y}_t$ is the expected power and y_t is the actual power, all at the same instance t. The residuals for each day and inverter are then grouped and summed.

The second step, finding the limits, is done by measuring the standard deviation of the daily residuals for each inverter. We set a daily lower limit (DLL) at three standard deviations and a daily upper limit (DUL) at five standard deviations, as described by equations 3 and 4. This simple method is inspired by (De Benedetti et al., 2018) and was more recently employed in (Syamsuddin et al., 2024).

$$DLL = 3\sigma_{daily} \tag{3}$$

$$DUL = 5\sigma_{daily} \tag{4}$$

The final step is simple: for each day when the residual value exceeds the DUL, we raise an alarm (value of 1); each time the DLL is crossed (but not the DUL), we raise a warning (value of 0.5).

5.5. Model evaluation metrics

The performance of the models is evaluated using two complementary criteria: (i) their ability to accurately model the inverter power output and (ii) their ability to detect anomalies. To quantify the prediction accuracy, two widely used regression metrics are employed: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) (Kim, Kim, Kim, Lee, & Yoon, 2025):

The second evaluation criterion assesses the anomaly detection capability of the proposed approach. In this context, the problem is formulated as a binary classification task where the system state is classified as either normal or abnormal. To evaluate the detection performance, the metrics precision, recall, and F1-score are used.

6. RESULTS

Table 4 summarises the forecasting and anomaly detection performance obtained by the different models under the various categorical feature encoding strategies. Two tasks are evaluated: the models' ability to accurately forecast inverter power output, measured by MAE and RMSE, and the effectiveness of the anomaly detection method, measured by precision, recall, and F1 score.

6.1. Power forecasting results

Across all evaluated models, including categorical inverter alarm variables, does not substantially improve short-term power-forecasting accuracy. In most configurations, models trained exclusively on continuous operational variables achieve performance comparable to or slightly better than those incorporating categorical alarm features. This suggests that additional work is required to obtain value from the categorical signals generated by the inverter. However, the best-performing algorithm in the power estimation task is the MLP with entity embeddings, pointing to an interesting line of investigation.

For the MLP model, entity embeddings yield the best forecasting results among the tested configurations, achieving the lowest MAE. However, the improvement compared to the model using only continuous variables is marginal. In contrast, one-hot encoding slightly degrades performance due to the high dimensionality and sparsity it introduces. Entity embeddings mitigate this issue by learning compact representations of categorical values.

The XGBoost model shows very similar forecasting performance with and without categorical variables. This suggests that tree-based models handle high-dimensional sparse inputs more effectively, while entity embeddings do not provide a significant advantage.

For the LSTM model, including categorical variables gener-

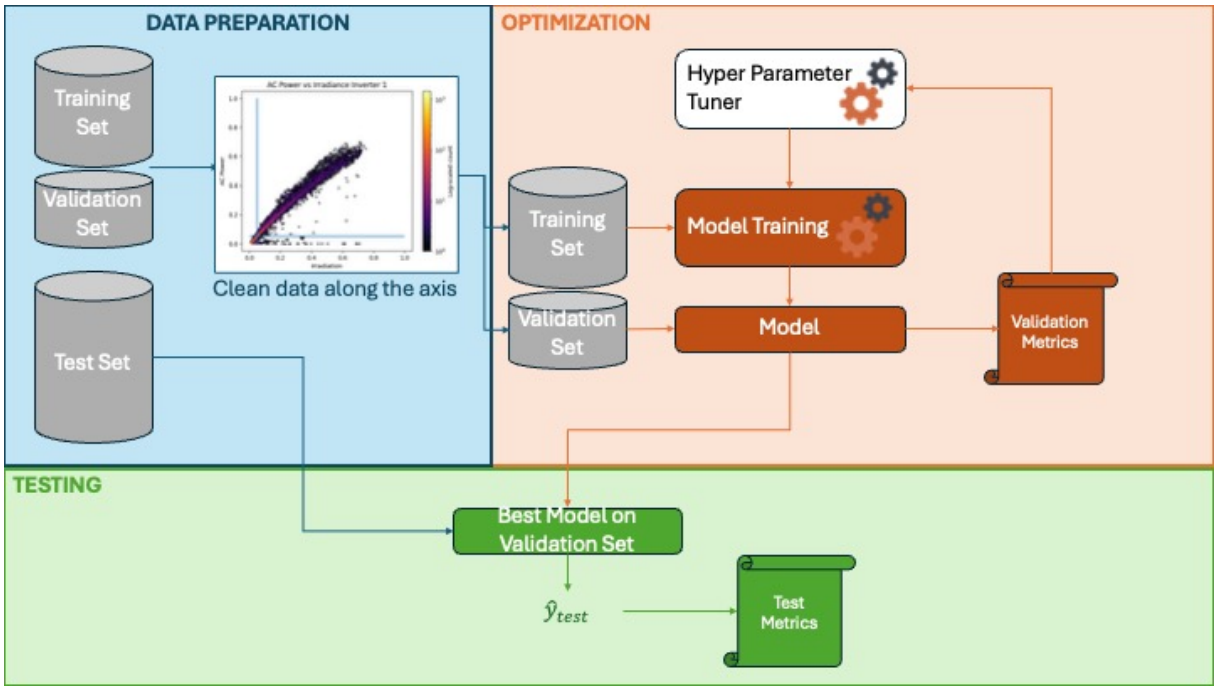

Figure 4. Data split, optimization process, and testing

ally leads to worse forecasting performance. This degradation is most pronounced with one-hot encoding, while entity embeddings partially mitigate the effect but still perform worse than using only continuous variables.

Overall, the best forecasting performance is achieved by the MLP model with entity embeddings, closely followed by the MLP model using only continuous variables. The strong performance of this relatively simple architecture suggests that long temporal dependencies play a limited role in predicting PV inverter power output. This observation is supported by the partial autocorrelation analysis (Figure 5), which shows that correlations quickly diminish beyond the most recent lag.

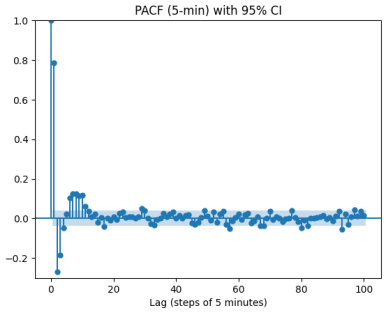

Figure 5. Partial auto-correlation function of the power output of an inverter

The MAE distributions across all inverters can be seen in Figure 6.

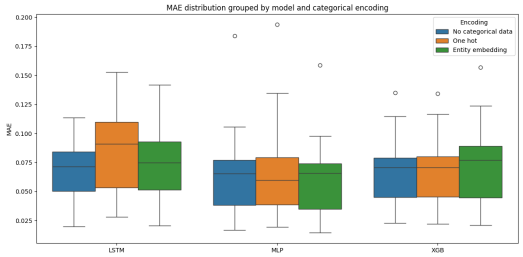

Figure 6. Box plot of MAE of the power prediction of the different models with different categorical data as input.

6.2. Anomaly detection performance

In the anomaly detection task, the inclusion of categorical variables does not consistently improve performance across the evaluated models. Precision values are generally high, while recall remains moderate, largely due to the strong class imbalance in the dataset, where PV systems operate under normal conditions most of the time. In addition, forecasting models are trained primarily on normal operating data, limiting their ability to learn patterns associated with abnormal behaviour.

For the MLP model, the best anomaly-detection performance is achieved when only continuous variables are used. Nevertheless, entity embeddings outperform one-hot encoding, again highlighting the advantage of dense learned representations over sparse categorical encodings.

The XGBoost model shows similar anomaly-detection per-

formance across different encoding strategies, indicating that tree-based models are relatively robust to sparse input features. Interestingly, the configuration with entity embeddings, which produces larger forecasting errors, achieves the highest precision and F1 score. This suggests that the model tends to overestimate expected power during abnormal conditions, resulting in larger residuals that facilitate anomaly detection.

For the LSTM model, using entity embeddings improves performance over one-hot encoding, although the overall anomaly-detection capability remains comparable to that of the MLP model.

An example of an underperforming inverter can be seen in Figure 7.

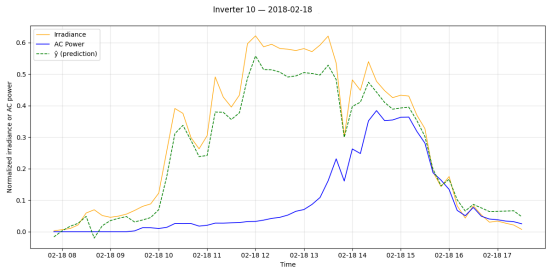

Figure 7. Underperformance case detected by the anomaly detection framework.

Although the inclusion of categorical alarm signals does not significantly improve the predictive performance of the evaluated models, the alarm distribution seen in Figure 2 suggests that these signals still contain valuable operational information. Certain alarm codes occur more frequently during abnormal operating periods, suggesting that categorical inverter messages may offer useful diagnostic insights. These observations suggest that categorical alarm data may be more suitable for complementary monitoring tasks, such as fault interpretation or event pattern analysis, rather than for directly improving the accuracy of short-term power forecasting.

7. CONCLUSION

This work investigates the potential value of categorical inverter alarm data in data-driven monitoring models for PV systems. To the best of the authors' knowledge, this study is the first to systematically explore the use of inverter alarm codes as categorical inputs in ML models designed for power prediction and anomaly detection in PV systems.

Several modelling approaches were evaluated to assess their ability to predict inverter power output and detect abnormal behaviour. The models compared include classical ML methods and DL architectures, namely MLP, XGBoost, and LSTM. The results show that relatively simple architectures provide competitive or superior performance compared to more complex models. In particular, the MLP model using only continuous operational variables achieved the best performance in the anomaly detection task, while the same model incorporating entity embeddings achieved the best results in the power prediction task.

Different encoding strategies for categorical alarm signals were also evaluated to analyse how inverter alarms can be integrated into ML models. Entity embeddings generally provide better results than one-hot encoding, highlighting the advantage of dense learned representations when handling categorical operational data. However, the inclusion of categorical alarm variables does not consistently improve power-forecasting accuracy or anomaly-detection performance. This suggests that inverter alarm signals contain limited predictive information for short-term power forecasting when used directly as model inputs.

Nevertheless, the exploratory analysis of alarm distributions reveals that certain error codes occur more frequently during abnormal operating periods (see Figure 2). This observation indicates that categorical inverter messages contain useful diagnostic information about system behaviour. Importantly, these signals are already generated internally by the inverter monitoring system and therefore do not require additional sensors or instrumentation, making them a potentially valuable source of information for industrial monitoring applications.

Future work will focus on several directions. First, improving the interpretability of alarm codes would allow a more informed selection of categorical inputs. In the present work, the semantic meaning of the numeric alarm codes was not available, and all codes were therefore treated equally. Access to detailed alarm descriptions could allow the identification of the most informative signals and improve the integration of categorical data into monitoring models.

Second, larger datasets containing more failure events would allow models to learn richer representations of rare alarm patterns and better capture relationships between alarm signals and abnormal operating conditions. Finally, future research will investigate more advanced approaches to exploiting categorical operational data, such as sequential pattern analysis of alarm events and other data-mining techniques to identify recurring fault signatures in inverter alarm logs.

This work establishes a baseline for integrating categorical inverter alarm data into PV monitoring models and highlights the need for further research into methods that can effectively exploit this largely untapped source of operational information.

Table 4. MAE and RMSE for different models and input configuration. Precision, recall, and F1 score of anomaly detection.

Model	Input Type	Avg. MAE	Avg. RMSE	Precision	Recall	F1
MLP	Without discrete	0.058 ±0.026	0.085 ±0.033	**0.932**	**0.593**	**0.706**
	One-Hot Encoded	0.062 ±0.028	0.090 ±0.040	0.874	0.544	0.648
	Entity Embedding	**0.057** ±0.024	**0.083** ±0.031	0.915	0.577	0.686
XGBoost	Without discrete	0.065 ±0.028	0.090 ±0.032	0.824	0.491	0.600
	One-Hot Encoded	0.066 ±0.025	0.090 ±0.030	0.838	0.514	0.623
	Entity Embedding	0.070 ±0.029	0.094 ±0.034	0.869	0.519	0.639
LSTM	Without discrete	0.069 ±0.023	0.094 ±0.027	0.881	0.546	0.646
	One-Hot Encoded	0.085 ±0.033	0.113 ±0.039	0.879	0.590	0.673
	Entity Embedding	0.074 ±0.028	0.099 ±0.032	0.899	0.566	0.664

REFERENCES

Bezerra, A., Silva, I., Guedes, L. A., Silva, D., Leitão, G., & Saito, K. (2019). Extracting Value from Industrial Alarms and Events: A Data-Driven Approach Based on Exploratory Data Analysis. *Sensors*, *19*(12), 2772. doi: 10.3390/s19122772

Chang, Z., & Han, T. (2024). Prognostics and health management of photovoltaic systems based on deep learning: A state-of-the-art review and future perspectives. *Renewable and Sustainable Energy Reviews*, *205*, 114861. doi: 10.1016/j.rser.2024.114861

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). doi: 10.1145/2939672.2939785

De Benedetti, M., Leonardi, F., Messina, F., Santoro, C., & Vasilakos, A. (2018). Anomaly detection and predictive maintenance for photovoltaic systems. *Neurocomputing*, *310*, 59–68. doi: 10.1016/j.neucom.2018.05.017

Guo, C., & Berkhahn, F. (2016). *Entity Embeddings of Categorical Variables.* arXiv. doi: 10.48550/arXiv.1604.06737

Gutschi, C., Furian, N., Suschnigg, J., Neubacher, D., & Voessner, S. (2019). Log-based predictive maintenance in discrete parts manufacturing. *Procedia CIRP*, *79*, 528–533. doi: 10.1016/j.procir.2019.02.098

Hashemi, B., Taheri, S., Cretu, A.-M., & Pouresmaeil, E. (2021). Systematic photovoltaic system power losses calculation and modeling using computational intelligence techniques. *Applied Energy*, *284*, 116396. doi: 10.1016/j.apenergy.2020.116396

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Ibrahim, M., Alsheikh, A., Awaysheh, F., & Alshehri, M. (2022). Machine Learning Schemes for Anomaly Detection in Solar Power Plants. *Energies*, *15*(3), 1082. doi: 10.3390/en15031082

International Energy Agency. (2025). *World Energy Outlook 2025* (Tech. Rep.).

Kim, J., Kim, H., Kim, H., Lee, D., & Yoon, S. (2025). A comprehensive survey of deep learning for time series forecasting: architectural diversity and open challenges. *Artificial Intelligence Review*, *58*(7), 216. doi: 10.1007/s10462-025-11223-9

Liebermann, S., Um, J.-S., Hwang, Y., & Schlüter, S. (2021). Performance Evaluation of Neural Network-Based Short-Term Solar Irradiation Forecasts. *Energies*, *14*(11), 3030. doi: 10.3390/en14113030

Luo, M., Li, X., Zhang, D., Zhao, Y., & Lim, P. (2008). Categorical data analysis for equipment failure prediction. In *2008 34th Annual Conference of IEEE Industrial Electronics* (pp. 1473–1478). IEEE. doi: 10.1109/IECON.2008.4758171

Marangis, D., Livera, A., Tziolis, G., Makrides, G., Kyprianou, A., & Georghiou, G. E. (2024). Trend-Based Predictive Maintenance and Fault Detection Analytics for Photovoltaic Power Plants. *Solar RRL*, *8*(24), 2400473. doi: 10.1002/solr.202400473

Onal, Y. (2022). Gaussian Kernel Based SVR Model for Short-Term Photovoltaic MPP Power Prediction. *Computer Systems Science and Engineering*, *41*(1), 141–156. doi: 10.32604/csse.2022.020367

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. doi: 10.1038/323533a0

Sheppard, S., Dickey, K. A., Koskey, S., Teasley, C., Perullo, C., Fregosi, D., & Li, W. (2024). Benchmarking a Physics-Based Approach for Anomaly Detection at Utility PV Plants. In *2024 IEEE 52nd Photovoltaic Specialist Conference (PVSC)* (pp. 0856–0858). IEEE. doi: 10.1109/PVSC57443.2024.10749158

Syamsuddin, A., Adhi, A. C., Kusumawardhani, A., Prahasto, T., & Widodo, A. (2024). Predictive maintenance based on anomaly detection in photovoltaic system using SCADA data and machine learning. *Results in Engineering*, *24*, 103589. doi: 10.1016/j.rineng.2024.103589

Yao, S., Kang, Q., Zhou, M., Abusorrah, A., & Al-Turki, Y. (2021). Intelligent and Data-Driven Fault Detection of Photovoltaic Plants. *Processes*, *9*(10), 1711. doi: 10.3390/pr9101711