Bridging AI and Industry: Research and Innovation from the CINI-AIIS Lab at Federico II University

Flora Amato[1], David Carlini[2], Alessandro Del Prete[1], Sofia Dutto[1], Antonio Galli[1,*], Narendra Patwardhan[1], Stefano Marrone[1], Vincenzo Moscato[1], Gabriele Piantadosi[2], Carlo Sansone[1] and Marco Valle[2]

[1]*University of Naples Federico II, Via Claudio 21, Naples, 80125, Italy*
[2]*Simar Group S.r.l., Viale I Maggio, 8, 63813 Zona Artigianale Cam FM*
[3]*ENEA, Centro Ricerche Portici, 80055 Portici (NA), Italy*

Abstract

Artificial intelligence (AI) is increasingly driving transformation across industrial domains, unlocking new levels of automation, precision, and operational intelligence. Leveraging Machine Learning (ML) and Deep Learning (DL), modern industrial systems can anticipate failures, recognize anomalies, and process visual data with high accuracy. Predictive maintenance benefits from ML models capable of analyzing operational patterns to optimize service intervals. In parallel, DL architectures such as Convolutional Neural Networks (CNNs) enhance visual inspection tasks, enabling effective detection of defects and monitoring of production quality. The synergy between DL and natural language processing (NLP) also supports automation in areas like document classification and inventory tracking. At the CINI-AIIS Lab of the University of Naples Federico II, several initiatives are exploring these AI-driven approaches to promote innovation and resilience within industrial applications.

Keywords

Predictive Maintenance, Energy Forecasting, Anomaly Detection, Remaining Useful Life, Voice Assistant

1. Introduction

Artificial Intelligence (AI) is rapidly becoming a foundational technology across a wide spectrum of industrial domains, enabling unprecedented levels of automation, adaptability, and data-driven decision making. By emulating certain aspects of human reasoning, AI offers solutions capable of addressing highly complex tasks that were traditionally managed through manual intervention or rule-based systems. In particular, sectors such as manufacturing, energy, and logistics are witnessing a significant transformation, driven by the integration of intelligent algorithms into core operational workflows. In the industrial context, AI plays a crucial role in enhancing productivity and operational resilience. Machine Learning (ML), a primary branch of AI, empowers systems to learn from historical data and adapt to dynamic conditions. This learning capacity is especially valuable in predictive maintenance, where algorithms analyze sensor data to estimate the optimal timing for equipment servicing, reducing the likelihood of unexpected failures and extending the operational lifespan of machinery. Similarly, ML techniques are instrumental in identifying subtle patterns of deviation in real-time, supporting anomaly detection across production lines, power grids, and renewable energy systems.

Deep Learning (DL), a subfield of ML, further amplifies the potential of AI by enabling systems to interpret complex, high-dimensional data. Convolutional Neural Networks (CNNs), for example, have shown remarkable effectiveness in industrial visual inspection, where they are employed for detecting surface defects, misalignments, or quality deviations in manufactured products. Beyond image processing, the integration of DL with Natural Language Processing (NLP) is streamlining administrative

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy
*Corresponding author.
flora.amato@unina.it (F. Amato); d.carlini@outlook.com (D. Carlini); alessandro.delprete@unina.it (A. D. Prete); sofia.dutto@unina.it (S. Dutto); antonio.galli@unina.it (A. Galli); narendraprakash.patwardhan@unina.it (N. Patwardhan); stefano.marrone@unina.it (S. Marrone); vincenzo.moscato@unina.it (V. Moscato); carlo.sansone@unina.it (C. Sansone); m.valle@simargroupsrl.com (M. Valle)
0000-0002-4807-5664 (N. Patwardhan); 0000-0001-6852-0377 (S. Marrone); 0000-0002-8176-6950 (C. Sansone)

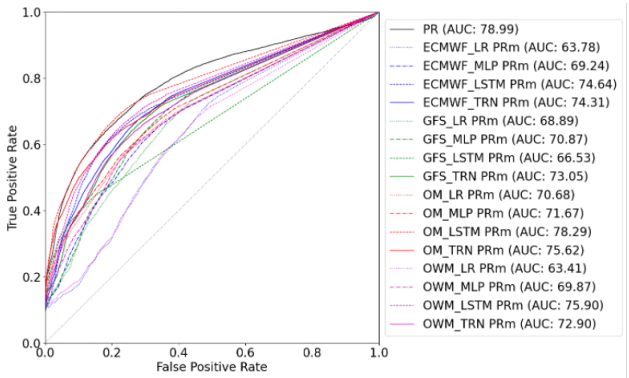

Figure 1: ROC curves illustrating the classification performance of the Performance Ratio (PR) and multiple machine learning models (Linear Regression, MLP, LSTM, Transformer) trained on meteorological inputs from different weather data providers (ECMWF, GFS, OWM, OM). The Area Under the Curve (AUC) is reported for each provider-model combination.

and logistic processes, such as automated analysis of technical documentation, intelligent inventory tracking, and parsing of unstructured industrial records. These technological advances are not confined to research labs they are being progressively adopted in real-world scenarios, particularly in energy systems (e.g., photovoltaic plant monitoring), smart manufacturing, and supply chain optimization. AI's ability to extract actionable insights from multimodal data sources is enabling industries to move from reactive strategies to proactive and autonomous systems.

In this paper, we present a set of applied research activities conducted by the CINI-AIIS Laboratory at the University of Naples Federico II, highlighting their impact on different industrial sectors. The initiatives discussed span from intelligent monitoring in energy production to deep learning-based quality control, with the shared objective of demonstrating how AI can serve as a catalyst for sustainable innovation and digital transformation in industry.

2. ML-based Anomaly detection for Photovoltaic systems

Photovoltaic (PV) systems, despite their growing role in the energy transition, remain susceptible to faults affecting modules, cabling, inverters, and other components. These issues often arise from environmental and electrical conditions, such as soiling, weather extremes, or equipment degradation [1]. Undetected, they can cause energy losses and increased costs, threatening long-term system efficiency. Therefore, accurate monitoring and timely fault detection are crucial to reduce downtime and maintain stable energy output [2]. Fault detection methods vary by input data, analytical techniques, and anomaly resolution. They are generally divided into signal-level and module-level approaches [3]. Signal-level methods analyze inverter data (e.g., voltage, current, power), while module-level techniques detect faults in individual panels using visual or thermal imaging, often acquired via drones or fixed cameras.

Within signal-level approaches, the Performance Ratio (PR), a metric standardized by IEC 61724-1 [4], is widely adopted for assessing PV system efficiency. It measures the ratio between actual and expected energy output - based on irradiance - and expresses it as a percentage. While it is computationally simple and requires minimal input data, it does not account for temperature effects or component degradation. Despite these limitations, the PR remains a valuable baseline for performance monitoring and anomaly detection. Our research examines the use of PR for fault detection across five PV systems, developing a detection framework based on performance metrics and KPIs. As presented in Figure 1, the PR-based model achieves promising results (AUC 80%, accuracy 78%), though a lower F1-score (54%) suggests room for improvement in detection precision and generalization.

We explored machine learning (ML) models as virtual sensors for fault detection and forecasting, removing the need for local sensors, which are often unreliable. These models leverage indicators such

as prediction error and model-derived metrics. The most effective, an LSTM trained on OpenMeteo data, achieved performance comparable to the PR-based approach (Figure 1). While the PR remains a useful baseline, ML techniques offer greater analytical depth and robustness, enabling more accurate diagnostics and proactive system optimization. Despite recent progress, AI methods face limitations due to scarce labeled data and the restricted sharing of sensitive production information. To overcome these issues, we explore two strategies: generating synthetic PV data to augment datasets and applying federated learning to enable privacy-preserving, decentralized model training.

Recently, Federated Learning (FL) has emerged as a promising approach to fault detection and performance prediction in PV systems, enabling collaborative pattern formation across distributed sites without sharing sensitive data. In this context, we are actively working on integrating FL techniques into our framework to improve detection capabilities while preserving data privacy and minimizing communication overhead.

3. Conditional TimeGANs for Photovoltaic Data Augmentation

One of the major limitations in the development of reliable AI-based solutions for photovoltaic (PV) energy forecasting and anomaly detection lies in the limited availability of large-scale, high-quality, and diverse datasets. PV energy production is inherently affected by numerous factors such as weather variability, geographic location, system configuration, and operational conditions. Capturing this variability in real datasets is often challenging due to cost constraints, data privacy issues, and the infrequent occurrence of rare or faulty operational scenarios. As a result, machine learning models trained on such limited datasets often suffer from poor generalization, reduced robustness in deployment, and inadequate performance when exposed to out-of-distribution or edge cases.

To address data scarcity in PV systems, recent research has focused on generative models capable of producing realistic time-series data. Among these, Conditional TimeGANs have proven particularly effective. Unlike traditional GANs, they incorporate temporal dynamics and use conditioning inputs—such as panel type, system capacity, irradiance, temperature, and humidity—to guide the data generation process. This enables the synthesis of scenario-specific sequences that better reflect real operational and environmental conditions.

The architecture combines a conditional GAN operating in latent space with a time-aware encoder-decoder (TED) that maps real PV and weather data in and out of this space. Training involves three loss functions: adversarial (for realistic outputs), reconstruction (to preserve temporal structure), and embedding (to enforce semantic similarity in the latent space). This design enables the generation of time series that are both realistic and contextually consistent.

Conditional TimeGANs offer a notable improvement over traditional data augmentation methods by effectively capturing complex and long-range temporal patterns. This makes them particularly suited for PV energy forecasting, where production is influenced by seasonal and hardware-related factors. They generate realistic synthetic data, even for rare scenarios, improving forecasting accuracy by up to 15% and fault detection by 20%. The data produced is both statistically valid and diverse, supporting the training of advanced models like transformers. Additionally, TimeGANs reduce the need for costly data collection and enable virtual testing of PV configurations, including fault conditions rarely observed in real-world datasets.

Their flexibility further allows integration with real-time data and domain-specific attributes, supporting hybrid training and domain adaptation across different regions or system types. This enables scalable AI deployment without the need for extensive local retraining, making Conditional TimeGANs a powerful tool for advancing sustainable and generalizable energy forecasting solutions. In summary, Conditional TimeGANs represent a powerful and flexible tool for the generation of realistic, temporally consistent, and context-aware synthetic data in the PV domain. Their application supports a wide range of tasks including short-term forecasting, long-term energy yield estimation, fault detection, scenario simulation, and predictive maintenance. By enabling more resilient and scalable AI models, they contribute to improving the efficiency and reliability of photovoltaic systems and facilitate broader

adoption of renewable energy technologies in line with global sustainability goals.

4. Enhancing On-Device Voice Assistant Intent Recognition via Hybrid LLM-Grammar Constrained Decoding

The efficacy of voice assistant systems critically depends on accurate text-to-intent recognition, the process of translating a user's natural language command into a structured, machine-interpretable format. This task is complicated by the inherent ambiguity and variability of human language, requiring a robust mechanism to map diverse phrasings to a canonical set of actions and associated parameters (e.g., converting "Dim the lights in the study to half" into `{"action": "set_brightness", "location": "study", "level": "50%"}`). In this work we present a hybrid approach that combines human expertise, Large Language Models (LLMs) of the Llama family, and formal grammar constraints to achieve robust intent recognition for on-device voice assistants. By integrating a context-free grammar (CFG) during decoding in *llama.cpp*, we strictly enforce syntactic correctness of model outputs. A human-in-the-loop workflow iteratively defines new intents, validates generated schemas, and augments the grammar—enabling compact models (e.g., 8B or even 1B parameters) to operate with high reliability and minimal computational overhead.

4.1. Grammar-Constrained Decoding

Grammar-constrained decoding formally ensures that an LLM's output conforms to a specified formal language, typically defined by a context-free grammar (CFG)[5, 6]. Let G be such a grammar whose language is $L(G)$. At each generation step i, given the already-generated token prefix $w_{<i}$, the objective is to identify the set of allowed next tokens, $A(w_{<i})$, such that any token $t \in A(w_{<i})$ can lead to a complete string in $L(G)$. Formally, the allowed-token set is:

$$A(w_{<i}) \;=\; \{t \in V_{\mathrm{LM}} \,:\, \exists s \in L(G) \text{ with prefix } w_{<i}t\}$$

where V_{LM} is the LLM's token vocabulary. Constrained decoding then modifies the next-token probability distribution, $p(t \mid w_{<i})$, by setting the probability of tokens not in $A(w_{<i})$ to zero. If ℓ_t represents the logit for token t produced by the LLM, the modified distribution $p'(t \mid w_{<i})$ becomes:

$$p'(t \mid w_{<i}) \;=\; \begin{cases} \dfrac{\exp(\ell_t)}{\sum_{u \in A(w_{<i})} \exp(\ell_u)}, & t \in A(w_{<i}), \\ 0, & \text{otherwise.} \end{cases}$$

In practice, this is often implemented by setting $\ell_t \to -\infty$ for all $t \notin A(w_{<i})$ before the softmax operation, guaranteeing that all generated outputs strictly adhere to G. This approach provides a strong guarantee of syntactic correctness, crucial for reliable intent parsing.

For the practical implementation of grammar-constrained decoding, we utilize the `llama.cpp` library. This framework is well-suited for our objectives due to its efficient C/C++ implementation, support for the Llama model architecture, and its integrated functionality for grammar-based control using a BNF-like notation termed GBNF. Within `llama.cpp`, a user-supplied GBNF grammar is compiled into an internal parser. During token generation, this parser interacts with the LLM's decoding loop, employing an "opportunistic masking" strategy. The LLM first proposes a token; if this token is syntactically valid according to the current grammar state, it is accepted. If invalid, `llama.cpp` consults the grammar to determine the full set of permissible next tokens and re-samples from this restricted set[7].

4.2. Hybrid Intent Definition and Refinement Methodology

We propose a hybrid, human-in-the-loop methodology for defining and refining voice assistant intents. The process initiates with a user (e.g., a system developer or domain expert) providing a natural language description of a desired new intent (e.g., "A user should be able to schedule a

meeting with specific attendees for a given date and time"). An LLM is then employed to parse this description and propose an initial structured representation, typically an action and a set of parameters, for example: `{action: "schedule_meeting", parameters: ["attendees", "date", "time", "subject"]}`. This machine-generated schema undergoes review and validation by the human operator, who may confirm or modify it to ensure semantic accuracy and completeness.

4.3. Example Generation and Grammar Integration

After validating the intent structure, the system uses an LLM to generate positive and negative examples. Positive examples match the intended action (e.g., "Book a meeting..."), while negative ones are similar but irrelevant or ambiguous (e.g., "What meetings do I have..."). All examples are reviewed and refined by humans. Subsequently, a templated generation component processes the validated intent schema and its associated examples. This component performs two critical functions:

1. It incorporates the natural language description of the intent [8], along with the curated positive and negative examples, into the system prompt provided to the LLM during runtime inference. This serves as rich, few-shot contextual guidance.
2. It dynamically updates the GBNF grammar file by generating and integrating new production rules that specifically define the syntactic structure for the newly added intent. This ensures that, at runtime, the LLM's output for this intent is strictly constrained to the defined `{action, parameters}` format.

4.4. Results and On-Device Deployment Implications

The proposed hybrid methodology—combining LLM scaffolding, human-in-the-loop validation, and grammar-constrained decoding—enables the development of reliable text-to-intent classifiers. Empirical results show improved accuracy and predictability over methods based solely on prompt engineering or fine-tuning large, unconstrained models. Grammar constraints also allow for deploying smaller models on-device: for example, a Llama 3.1 8B with GBNF constraints can match the performance of much larger models. This opens the door to using even smaller versions, like Llama 3.2 1B, for efficient, low-latency intent recognition, enhancing responsiveness in voice assistant applications.

5. RUL Prediction: A Performance Comparison of LSTM and Transformer

Entering the realm of predictive maintenance means moving from fixing things after they break to anticipating problems before they occur [9]. This is a crucial change in industries where the smooth running of equipment is not only about efficiency but often also about safety and cost efficiency. A key challenge is to accurately predict the remaining useful life (RUL) of vital components; in essence, we know how long something will still function reliably.

In our study, we really wanted to understand how different deep learning approaches handle this task of RUL prediction, especially when dealing with complex time series data such as those obtained from machinery monitoring [10]. We decided to compare two important types of neural networks: LSTM networks, which have long been used for sequential data because they are able to remember patterns over time, and the newer Transformer architectures, known for their powerful attention mechanisms that can detect connections over long sequences. As a testing ground, we used NASA's popular C-MAPSS dataset, which simulates the degradation of turbofan aircraft engines over time, providing multivariate sensor data and operational parameters that reflect the engine's health status as it approaches failure. After rigorously implementing and evaluating both models, we examined their performance across several evaluation metrics: MAE, MSE, RMSE and the coefficient R^2.

The detailed results can be found in Table 5. The result is quite interesting: although the Transformer model seems to learn well during training, the LSTM model prevailed when we examined these final

evaluation metrics. This suggested to us that, perhaps, for datasets with certain characteristics - typically when the data is not too abundant or has a bit more noise - LSTMs could still offer more robust performance. That said, the potential of Transformers, particularly with large volumes of data, is undeniable and certainly deserves to be explored further.

	MAE	MSE	RMSE	R^2
LSTM	20.5	922.8	30.36	0.73
Transformer	29.1	1609.3	40.11	0.52

Table 1
Comparison of metric results for both models

Overall, this area of predictive maintenance, especially when using advanced artificial intelligence techniques, still has a lot of ground to cover. There is exciting work to be done in fine-tuning these models, perhaps combining their strengths in hybrid approaches and finding practical ways to deal with real-world obstacles such as limited data or computational demands. The goal, of course, is to make predictive maintenance an even more powerful and reliable tool in all areas.

Acknowledgments

This work was supported in part by the Piano Nazionale Ripresa Resilienza (PNRR) Ministero dell'Università e della Ricerca (MUR) Project under Grant PE0000013-FAIR

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] P. Branco, F. Gonçalves, A. C. Costa, Tailored algorithms for anomaly detection in photovoltaic systems, Energies 13 (2020) 225. doi:https://doi.org/10.3390/en13010225.

[2] A. Mellit, G. Tina, S. Kalogirou, Fault detection and diagnosis methods for photovoltaic systems: A review, Renewable and Sustainable Energy Reviews 91 (2018) 1–17. doi:https://doi.org/10.1016/j.rser.2018.03.062.

[3] G. M. El-Banby, N. M. Moawad, B. A. Abouzalm, W. F. Abouzaid, E. Ramadan, Photovoltaic system fault detection techniques: a review, Neural Computing and Applications 35 (2023) 24829–24842. doi:https://doi.org/10.1007/s00521-023-09041-7.

[4] International Electrotechnical Commission, IEC 61724-1:2021 – Photovoltaic system performance – Part 1: Monitoring, https://webstore.iec.ch/en/publication/65561, 2021. Second edition, International standard, Geneva, Switzerland.

[5] S. Geng, M. Josifoski, M. Peyrard, R. West, Grammar-constrained decoding for structured nlp tasks without finetuning, arXiv preprint arXiv:2305.13971 (2023).

[6] K. Park, T. Zhou, L. D'Antoni, Flexible and efficient grammar-constrained decoding, arXiv preprint arXiv:2502.05111 (2025).

[7] L. Beurer-Kellner, M. Fischer, M. Vechev, Guiding llms the right way: Fast, non-invasive constrained generation, arXiv preprint arXiv:2403.06988 (2024).

[8] B. Wang, Z. Wang, X. Wang, Y. Cao, R. A Saurous, Y. Kim, Grammar prompting for domain-specific language generation with large language models, Advances in Neural Information Processing Systems 36 (2023) 65030–65055.

[9] A. Meddaoui, M. Hain, A. Hachmoud, The benefits of predictive maintenance in manufacturing excellence: a case study to establish reliable methods for predicting failures, The International Journal of Advanced Manufacturing Technology 128 (2023) 3685–3690.
[10] W. Zhang, D. Yang, H. Wang, Data-driven methods for predictive maintenance of industrial equipment: A survey, IEEE systems journal 13 (2019) 2213–2227.