Using Retrieval-Augmented Generation for Fault Prediction and Maintenance on the Factory Floor

Adhir Soechit
Department of Computer Science
University of Liverpool
Liverpool, United Kingdom
a.k.soechit@liverpool.ac.uk

Patrick Hosein
Department of Electrical and Computer Engineering
The University of the West Indies
St. Augustine, Trinidad and Tobago
patrick.hosein@uwi.edu

***Abstract*—In the rapidly evolving landscape of Industry 4.0, predictive maintenance is crucial for maximizing operational uptime and efficiency. Traditional predictive maintenance approaches, often reliant on static models, face challenges in addressing novel anomalies and providing contextually rich, actionable insights. An innovative Retrieval-Augmented Generation (RAG) system is proposed, leveraging Large Language Models (LLMs) for real-time anomaly detection and predictive maintenance. The system dynamically integrates machine sensor data with repair manual documentation and worker insights to generate highly relevant and interpretable diagnostic maintenance recommendations which are passed along in real-time to relevant personnel. The methodology aims to demonstrate the RAG system's ability to enhance decision-making, optimize industrial operations and mitigate equipment failures in complex manufacturing environments.**

***Index Terms*—Retrieval-Augmented Generation, RAG, Large Language Models, LLMs, Industry 4.0, Predictive Maintenance, Anomaly Detection, Smart Factories**

I. Introduction

The Fourth Industrial Revolution, Industry 4.0, is profoundly impacting various industrial sectors through the pervasive integration of digital technologies [1]. This era, characterized by the convergence of the Internet of Things (IoT), Artificial Intelligence (AI) and cyber-physical systems, transforms traditional manufacturing into highly efficient, autonomous smart factories [2]. A critical enabler of this transformation is predictive maintenance, a strategy aimed at preventing equipment failures before they occur. By anticipating potential breakdowns, predictive maintenance maximizes operational uptime, reduces costly unplanned downtime, and optimizes asset utilization across complex industrial environments [3].

However, despite significant benefits, traditional predictive maintenance approaches often face inherent limitations. These methods, frequently reliant on statistical models or predefined rule-based systems applied to sensor data, can struggle with the identification of novel or complex anomalies [4]. They often necessitate extensive manual feature engineering and typically lack the sophisticated contextual reasoning capabilities required to provide nuanced diagnostics and truly actionable recommendations. A further challenge in advancing research in this domain is the proprietary nature and limited accessibility of real-world industrial operational data, which often hinders the development and rigorous evaluation of advanced AI solutions.

Advanced AI paradigms, particularly Large Language Models (LLMs), offer a promising avenue to overcome these limitations. LLMs possess remarkable capabilities in understanding complex language, synthesizing information, and generating human-interpretable text [5]. When augmented with external knowledge, these models can move beyond pattern recognition to provide deep contextual insights. Retrieval-Augmented Generation (RAG) systems represent a cutting-edge approach that combines the generative power of LLMs with contextual, up-to-date knowledge retrieved from external data sources [6]. This integration mitigates common LLM challenges such as hallucination and outdated information, enhancing their reliability for knowledge-intensive tasks.

To address the challenges in predictive maintenance, the application of RAG systems within an intelligent agent framework is explored. The objective is to design, implement and rigorously evaluate a RAG-based LLM system, integrated with specialized monitoring and communication agents, tailored for real-time anomaly detection and comprehensive predictive maintenance in Industry 4.0 settings. A key methodological aspect involves the use of synthetically generated industrial data, encompassing both realistic time-series sensor readings and comprehensive technical documentation. This approach ensures the reproducibility and verifiability of findings, providing a robust framework for future investigations in this domain. Through its agentic components, the proposed system integrates operational and contextual data to deliver relevant, interpretable, and actionable insights, enhancing industrial decision-making and bolstering manufacturing resilience.

II. Related Work

The rapid evolution of Industry 4.0 has introduced unprecedented opportunities for predictive maintenance by integrating cyber-physical systems, industrial IoT and AI. Central to this transformation is the fusion of real-time monitoring, contextual reasoning and intelligent decision-making: an intersection where LLMs, RAG and agent-based architectures converge to enable sophisticated, interpretable and responsive maintenance frameworks [1], [2].

Available Soon

Predictive maintenance (PdM) aims to preempt equipment failure through continuous monitoring and analysis of operational data. Conventional PdM methods rely heavily on statistical models, rule-based systems and time-series analysis to detect degradation or abnormal patterns in signals [3]. These methods, though reliable in stable environments, struggle to generalize to novel failure types, often requiring manual feature engineering and lack interpretability - issues that limit their deployment in complex manufacturing contexts [4].

Recent years have seen increased use of deep learning models, particularly Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNNs) and autoencoders, which are used to capture non-linear temporal dependencies in sensor data [5]. However, despite improved detection capabilities, these models operate in isolation from domain knowledge embedded in technical documentation or expert manuals. Consequently, they offer little in the way of reasoning or actionable insight beyond anomaly flags, which restricts their utility for real-time, human-interpretable decision making.

LLMs, like GPT-3 and PaLM, have revolutionized natural language processing (NLP) with their ability to perform zero-shot reasoning and generate fluent, human-like text across various domains [6]. Nevertheless, their reliance on static parametric knowledge introduces factual limitations and are prone to hallucinations - generating responses that are fluent but incorrect or unverifiable [7].

To mitigate these challenges, the RAG framework augments LLMs with a retrieval mechanism that dynamically fetches relevant context from an external knowledge base at inference time. This enables the model to ground its outputs in specific, verifiable information, thereby improving factual accuracy and traceability. RAG has proven effective in open-domain question answering [8], biomedical diagnostics [9] and software troubleshooting [10], yet its deployment in time-sensitive, real-time industrial applications remains underexplored.

Most current implementations of RAG are designed for static or batched input, where latency is less critical. This operational constraint limits their direct applicability to high-throughput environments such as smart manufacturing, where diagnostics must be both timely and contextually rich [11]. Moreover, integration of RAG-enabled LLMs with real-time sensor data and unstructured industrial knowledge bases poses unique challenges in terms of heterogeneous data types, synchronization and reasoning under uncertainty.

Agent-based models (ABMs) have been utilized in industrial settings for their modularity, autonomy and suitability for distributed environments. These agents are capable of sensing, reasoning and acting either independently or cooperatively [12]. In manufacturing, ABMs have been applied to decentralized scheduling, condition monitoring and fault diagnosis.

The recent paradigm of agentic LLMs builds upon this foundation by embedding LLMs within structured workflows, allowing them to interact with external tools, like APIs and data streams. These agents follow iterative planning and decision-making protocols, enabling them to solve complex multi-step tasks in dynamic environments [13]. Architectures like ReAct (Reason + Act) and AutoGPT exemplify this trend, where LLMs not only generate responses, but also execute actions, query knowledge sources and adapt based on feedback [14].

Research has shown the effectiveness of such agentic frameworks in areas including scientific research automation [15], web navigation [16] and code generation [17]. However, applications in PdM - especially those involving real-time decision support based on sensor analytics and contextual knowledge, are still in nascent stages.

To realize responsive, human-interpretable PdM, this research proposes an architecture with three tightly coordinated components: a Monitoring Agent, a RAG-enabled LLM Reasoning core, and a Communication Agent. Each has been explored independently in literature, but their combination into a unified real-time framework for PdM remains underdeveloped.

Monitoring agents typically employ lightweight ML models or rule-based systems to detect anomalies in sensor data [18]. While effective at identifying out-of-spec conditions, they lack the reasoning capabilities necessary for root-cause analysis or prescriptive guidance. Integrating them with LLMs that can query and synthesize from rich documentation bridges this gap, as the LLM can explain anomalies in light of procedural, contextual and historical information.

The RAG-enabled Reasoning Core serves as the diagnostic engine. By pairing retrieval mechanisms with generative capabilities, it constructs grounded, multi-source explanations. Hybrid diagnostic models have shown that combining symbolic reasoning and deep learning improves interpretability and reduces error rates in high-stakes environments [19]. The RAG setup also supports provenance tracking, allowing outputs to be traced back to their source passages - crucial for transparency in operational decision-making.

The Communication Agent operationalizes insights by aligning output messages with personnel roles, notification protocols, and urgency levels. Literature on human-centered AI emphasizes the importance of not just delivering correct information, but doing so in a form that is actionable and role-specific [20]. The Communication Agent ensures that intelligence is embedded within organizational workflows and is not limited to generating a technician alert, opening a ticket in a Computerized Maintenance Management System, or notifying a supervisor.

III. Methodology

This section provides an overview of the design, implementation and operational pipeline of the RAG system developed for real-time anomaly detection and predictive maintenance within Industry 4.0 environments. The proposed framework integrates sophisticated machine learning models with advanced natural language processing capabilities, orchestrated through an intelligent agent architecture. The methodology is grounded in the use of synthetically generated data to ensure the reproducibility and verifiability of the experimental findings.

A. System Architecture

The system is conceptualized as a multi-agent framework, operating cohesively to provide continuously monitoring, intelligent diagnostics and proactive communication (see Figure 1). The high-level architecture comprises of the following primary components:

- Data Ingestion Layer: This layer is responsible for the continuous simulation and feeding of multi-sensor time-series data, mimicking real-world industrial sensor streams from various machines.
- Monitoring Agent: This agent continuously analyzes incoming sensor data streams to identify deviations from established normal operational patterns using an LSTM-based forecasting model.
- RAG-enabled LLM Reasoning Core: Upon detection of a potential anomaly by the Monitoring Agent, this core is activated. It leverages a robust knowledge base to retrieve contextually relevant information and employs an LLM to synthesize this information, generating highly interpretable diagnostic explanations and actionable maintenance recommendations.
- Communication Agent: The final stage involves the Communication Agent, which is tasked with intelligent dissemination of the generated insights. It dispatches tailored notifications to predefined personnel, considering their roles, urgency levels, and established communication protocols.

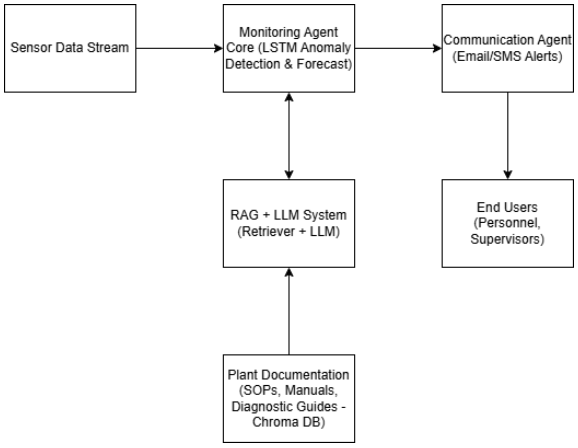

Fig. 1. AI-Driven Smart Manufacturing Monitoring System Architecture.

These agents operate asynchronously, facilitating a a responsive and near real-time decision-making loop essential for dynamic industrial settings. The underlying orchestration is managed through a LangGraph-based state machine, enabling complex multi-step reasoning and tool utilization.

This modular design enhances system resilience and scalability, allowing for independent development and deployment of each agent. The agentic approach enables dynamic decision-making and adaptive responses, which are crucial for navigating the complexities inherent in modern industrial operations and ensuring robust system performance.

B. Synthetic Data and Knowledge-Base Management

To ensure reproducibility, synthetic time-series sensor data and technical documentation are utilized. Data includes normal ranges and controlled anomalies. Sensor data (temperature, pressure, vibration and current) is loaded, cleaned and processed from CSVs. Technical documents (e.g. manuals) are segmented, embedded and indexed in a ChomaDB vector database, forming the RAG system's external knowledge source.

C. Anomaly Detection and Classification

The Monitoring Agent, implemented via a 'MonitoringAgentCore' class, handles data buffering, LSTM-based anomaly detection, and anomaly classification. Sensor readings for each machine are added to a fixed-size buffer (of sequence length 30). Raw data are normalized using a `RobustScaler`, fitted on training data. This method scales features robustly using the median and the interquartile range (IQR). The scaled value x_{scaled} for a raw value x_{raw} is calculated as:

$$x_{scaled} = \frac{x_{raw} - \text{median}}{\text{IQR}} \tag{1}$$

A sequential LSTM forecasting model, built with TensorFlow/Keras, learns normal temporal patterns to predict the next sensor values based on historical sequences. The model is trained to minimize the difference between its predictions and actual subsequent observations. Anomalies are detected based on the prediction error (Mean Squared Error, MSE) between the model's forecast for the next time step and the actual observed values. For a given time step t, and across D sensor dimensions (on normalized values $x'_{t,d}$ for actual and $\hat{x}'_{t,d}$ for predicted), the prediction error E_t is computed as

$$E_t = \frac{1}{D}\sum_{d=1}^{D}(x'_{t,d} - \hat{x}'_{t,d})^2 \tag{2}$$

where $x'_{t,d}$ is the actual normalized value at time t for sensor dimension d, and $\hat{x}'_{t,d}$ is the model's predicted normalized value for time t for sensor dimension d. If untrained, a simulated 'quick training' on dummy normal data sets a dynamic anomaly threshold, $E_{threshold}$ (e.g., the 99th percentile of training errors). An anomaly is flagged if E_t exceeds $E_{threshold}$. Classification also considers actual or predicted sensor values against predefined critical (`min`/`max`) or normal operating (`normal_low`/`normal_high`) thresholds. Deviations within thresholds are classified as 'unusual patterns (LSTM detected, but within defined thresholds)'.

D. RAG-enabled LLM Reasoning Core

This core provides intelligent diagnostics by combining LLM capabilities with external knowledge. Upon anomaly detection, a detailed natural language query is constructed, including machine identifier, timestamp, anomaly type, and the present sensor values. This query is used to semantically search the ChromaDB vector store, retrieving relevant passages from the synthetic technical documentation. The

retrieved context and the query from an augmented prompt, which is fed to the Llama 3.1 LLM. the LLM then generates a detailed explanation, present status, and actionable recommendations, grounding its response in the retrieved information to minimize hallucinations.

E. Communication Agent

The Communication Agent, using the `send_notification` tool, dispatches targeted and timely notifications. A contact personnel dictionary maps machine identifiers to specific roles (e.g., Specialist, Boss) and their contact information (email, phone). Upon receiving an anomaly and explanation, the `send_notification` tool is invoked asynchronously. Notifications include a clear subject line and a detailed body with machine, anomaly type, sensor values, and recommended actions.

To prevent alert fatigue, a 300-second (5-minute) cool-down is implemented. If a notification for a machine was sent recently, subsequent alerts for that machine are suppressed until the cool-down period elapses.

IV. Results

The AI-driven smart manufacturing monitoring system was evaluated on a large, synthetically generated dataset representing 15,000 time-series observations per industrial machine. The test data included both baseline operating conditions and engineered anomalies, such as gradual parameter drifts and abrupt sensor spikes, across all monitored signals. These behaviours were incorporated to mimic real-life data as close as possible. The LSTM-based Monitoring Agent's predictive accuracy was assessed by comparing the model's next-step predictions to the actual sensor readings. Table I summarizes the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for each sensor on the test set pertaining to machine M1.

TABLE I
Sensor-wise Prediction Accuracy Metrics for Machine 1 (M1)

Sensor	MAE	RMSE
Temperature	2.4598	3.1080
Pressure	2.0717	2.6113
Vibration	1.3768	1.9648
Current	2.9979	4.0546

These results indicate that the model reliably tracks the general trends and abrupt shifts in sensor readings, with error magnitudes remaining within acceptable operational margins for most parameters. Prediction accuracy is highest for vibration and pressure, while current readings display slight greater variance, likely due to the engineered anomaly spikes.

The model's predictive performance can be visualized by plotting actual versus predicted sensor values over a representative time window. Figure 2 demonstrates the LSTM model's ability to track dynamic changes across all four sensor parameters, accurately reflecting baseline stability, gradual drifts and engineered anomaly events. This close alignment between the model's predictions and true measurements validates the efficacy of the LSTM forecasting core under both normal and abnormal conditions.

Upon identification of a critical anomaly (e.g. pressure drops below minimum operational threshold), the system successfully invokes the RAG-enabled LLM reasoning core to generate a contextualized, evidence-based diagnosis. The Communication Agent then delivered a notification to relevant personnel as demonstrated in Figure 3. This notification, equipped with a cool-down mechanism to avoid alert fatigue, ensures timely and actionable dissemination of critical insights to both operators and supervisory staff.

V. Discussion

The results highlight the system's capabilities in accurately identifying anomalies, providing contextually rich diagnostics, and effectively managing alert dissemination, thereby addressing several limitations inherent in traditional predictive maintenance approaches.

The high sensitivity observed in the Monitoring Agent's LSTM forecasting model, facilitated by the low anomaly threshold, suggests a significant potential for the early detection of subtle operation deviations. This proactive capability is paramount for minimizing unplanned downtime and optimizing operational efficiency, as it enables intervention before minor equipment issues escalate into critical failures. The system's capability to classify anomalies beyond simple threshold breach,es by recognizing unusual temporal patterns, represents a notable advancement over conventional rule-based systems that often lack the sophistication for such nuanced detection.

The performance of the RAG-enabled LLM Reasoning Core underscores the value of augmenting large language models with external, domain-specific knowledge. The consistent generation of contextually aligned, evidence-based, and actionable recommendations address a critical gap in many data-driven preventative maintenance solutions: the pervasive lack of interpretability and prescriptive guidance. By grounding the LLM's outputs in a verifiable knowledge base, the system effectively mitigates common challenges such as factual inaccuracies or "hallucinations", thereby increasing the trustworthiness of the diagnostic insights provided to human operators. This interpretability is paramount for effective decision-making in high-stakes industrial contexts.

Furthermore, the operational success of the Communication Agent, particularly the effective implementation of the notification cool-down mechanism, is a practical consideration for any real-time monitoring system. In continuous monitoring environments, preventing alert fatigue is essential for maintaining the responsiveness and effectiveness of maintenance personnel. The role-based notification system ensures that relevant information is directed to the appropriate stakeholders, potentially optimizing response times and resource allocation.

Despite these promising results, the inherent limitations stemming from the exclusive use of synthetically generated data necessitate further empirical validation in complex, noisy,

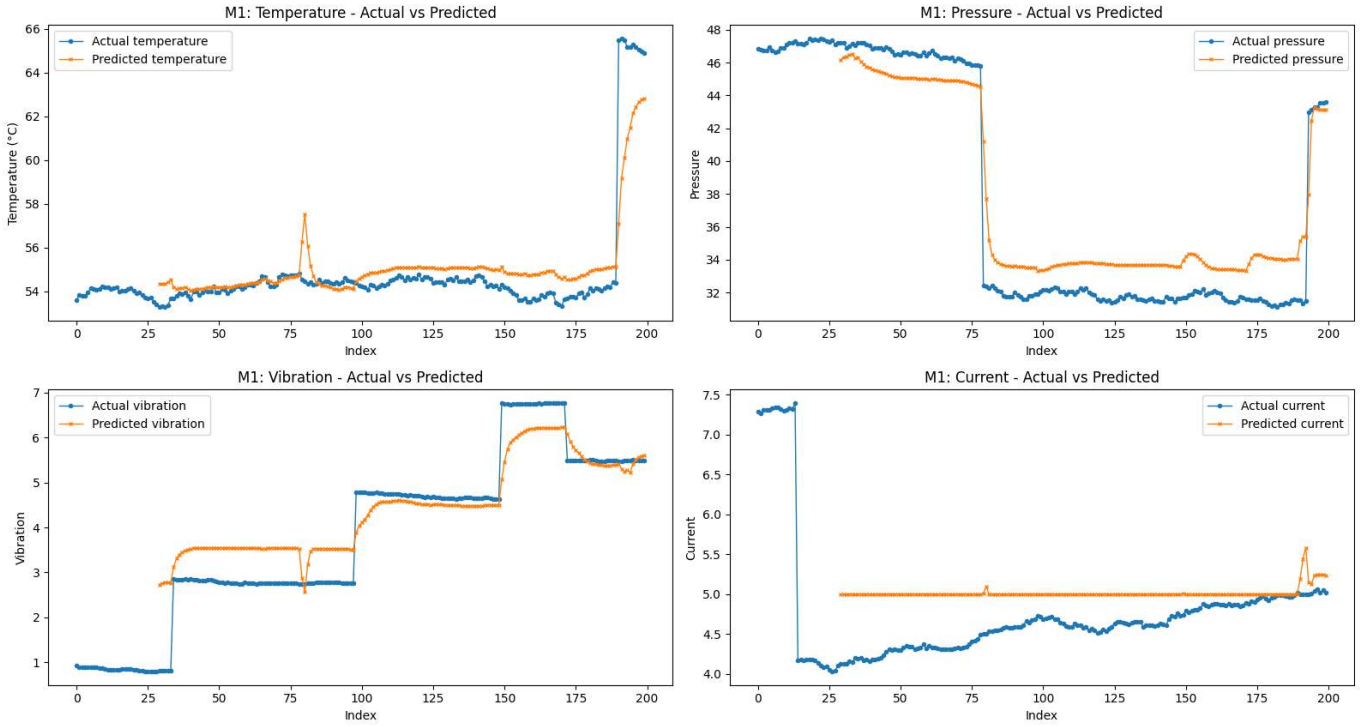

Fig. 2. Actual vs. predicted sensor values for machine M1 across all four monitored parameters (temperature, pressure, vibration, and current).

```
Notification sent to Senior Dr. Jones (Head of Engineering) (jones.eng@example.com, 555-0103):
Subject: Urgent Anomaly Alert: Machine M1 - CRITICAL: high vibration (2.69 > 2.00), high prediction error for vibration
Body: Machine: M1
Anomaly Type: CRITICAL: high vibration (2.69 > 2.00), high prediction error for vibration
Details and Recommended Action:
Based on the provided context, I can analyze the anomaly detected in Machine M1.

The machine is experiencing high vibration (2.69 mm/s, exceeding the threshold of 2.00) and a high prediction error for vibration. This indicates that the system has identified an issue with the servo motors or the robo
tic arm's alignment, as mentioned in "M1 Service Manual, Section 4.2 - Vibration Diagnostics".

To resolve this issue, refer to the relevant documentation:

1. Check motor mounts and bearing lubrication (Section 4.2).
2. Verify that the bearing lubrication is up-to-date and correctly applied.
3. Inspect the robotic arm's alignment and adjust it if necessary.

Monitor the machine's vibration levels and sensor readings after making these adjustments to ensure the issue has been resolved.

**Current Status:** The machine is currently experiencing a critical anomaly with high vibration, which could lead to equipment damage or product defects if not addressed promptly.

**Recommendation:** Stop Machine M1 immediately to prevent further damage. Perform the necessary checks and adjustments as outlined in "M1 Service Manual, Section 4.2 - Vibration Diagnostics".
```

Fig. 3. Example of a notification generated by the Communication Agent.

and unpredictable real-world industrial environments. This includes he need for extensive training of the LSTM forecasting model on diverse, real-world operational data and the expansion of the currently limited synthetic technical documentation to a comprehensive, continuously updated, and potentially heterogeneous knowledge base. Future work will therefore focus on integration the system with actual industrial sensor data streams and maintenance logs for robust empirical validation and performance benchmarking. This will be complemented by exploring more advanced LLMs and embedding models, investigating multi-sensor fusion techniques to enhance anomaly detection and diagnostic capabilities, and developing a user-friendly interface for real-time visualization. Ultimately, integrating the Communication Agent with existing Computerized Maintenance Management Systems (CMMS) aims to automate work order generation and streamline maintenance workflows, thereby enhancing the system's utility within a complete Industry 4.0 ecosystem.

VI. Conclusion

This research demonstrates the development and preliminary evaluation of an AI-driven smart manufacturing monitoring system, leveraging RAG with LLMs for real-time anomaly detection and predictive maintenance. The system's core components, a robust Monitoring Agent utilizing LSTM for anomaly detection, a RAG-enabled LLM Reasoning Core for contextual diagnostics, and an intelligent Communication Agent for targeted alerts, work cohesively to enhance opera-

tional efficiency and mitigate equipment failures. By integrating machine sensor data with a verifiable knowledge base, the system provides highly relevant, interpretable and actionable insights, addressing critical limitations of traditional predictive maintenance approaches and significantly improving decision-making in complex manufacturing environments.

The LSTM-based Monitoring Agent accurately identifies anomalies, from subtle deviations to critical breaches, by analyzing complex sensor data patterns. This proactive detection minimizes downtime and optimizes operational efficiency, preventing minor issues from escalating. The RAG-enabled LLM Reasoning Core transforms anomaly flags into rich, evidence-based diagnostics and actionable recommendations, enhancing trustworthiness and empowering personnel with contextual understanding for timely decisions. The Communication Agent's practical utility, with its role-based notifications and cool-down mechanism, ensures critical intelligence is disseminated efficiently, preventing alert fatigue and streamlining maintenance. This comprehensive approach, spanning detection, diagnosis and intelligent communication, marks a significant step towards autonomous and resilient smart factory operations.

Having demonstrated its core capabilities, the current system provides a foundation upon which future research can enhance its real-world applicability. This involves integrating with actual industrial sensor data and expanding the knowledge base beyond synthetic data for robust validation. The Communication Agent will evolve by connecting to HR databases for personnel rostering and interfacing with plant safety systems to trigger sirens or provide precise machine shutdown sequences during emergencies. Integrating with CMMS will automate work order generation, streamlining maintenance workflows. Developing user interfaces for real-time visualization will improve interaction. Continuous research into ethical AI and cyber security is paramount for responsible deployment. These advancements will ensure the system becomes more robust, intelligent and an indispensable tool for predictive maintenance and operational resilience in Industry 4.0.

REFERENCES

[1] J. Lee, B. Bagheri, and H.-A. Kao, "A cyber-physical systems architecture for industry 4.0-based manufacturing systems," *Manufacturing Letters*, vol. 3, pp. 18–23, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S221384631400025X

[2] W. Zhang, D. Yang, and H. Wang, "Data-driven methods for predictive maintenance of industrial equipment: A survey," *IEEE Systems Journal*, vol. 13, pp. 2213–2227, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:164483480

[3] P. Malhotra, L. Vig, G. M. Shroff, and P. Agarwal, "Long short term memory networks for anomaly detection in time series," in *The European Symposium on Artificial Neural Networks*, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:43680425

[4] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao, "Deep learning and its applications to machine health monitoring," *Mechanical Systems and Signal Processing*, vol. 115, pp. 213–237, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888327018303108

[5] J. Heaton, "Ian goodfellow, yoshua bengio, and aaron courville: Deep learning," *Genetic Programming and Evolvable Machines*, vol. 19, pp. 305–307, 2018. [Online]. Available: https://doi.org/10.1007/s10710-017-9314-z

[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, and et al., "Language models are few-shot learners," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[7] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, and et al., "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, Mar. 2023. [Online]. Available: https://doi.org/10.1145/3571730

[8] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, and et al., "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[9] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and et al., "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, pp. 1234–1240, 2 2020.

[10] M. R. Parvez, W. Ahmad, S. Chakraborty, B. Ray, and K.-W. Chang, "Retrieval augmented code generation and summarization," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2719–2734. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.232/

[11] V. Gungor and G. Hancke, "Industrial wireless sensor networks: Challenges, design principles, and technical approaches," *Industrial Electronics, IEEE Transactions on*, vol. 56, pp. 4258 – 4265, 11 2009.

[12] R. Calegari, G. Ciatto, V. Mascardi, and A. Omicini, "Logic-based technologies for multi-agent systems: a systematic literature review," *Autonomous Agents and Multi-Agent Systems*, vol. 35, p. 1, 2020. [Online]. Available: https://doi.org/10.1007/s10458-020-09478-3

[13] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, and K. e. a. Narasimhan, "React: Synergizing reasoning and acting in language models," in *International Conference on Learning Representations (ICLR)*, 2023.

[14] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, and et al., "Guidelines for human-ai interaction," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–13. [Online]. Available: https://doi.org/10.1145/3290605.3300233

[15] A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, and P. Schwaller, "Augmenting large language models with chemistry tools," *Nature Machine Intelligence*, vol. 6, pp. 525–535, 5 2024.

[16] S. Zhou, F. F. Xu, H. Zhu, X. Zhou, R. Lo, A. Sridhar, and et al., "Webarena: A realistic web environment for building autonomous agents," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=oKn9c6ytLx

[17] X. Wang, Y. Chen, L. Yuan, Y. Zhang, Y. Li, H. Peng, and et al., "Executable code actions elicit better llm agents," in *Proceedings of the 41st International Conference on Machine Learning*, ser. ICML'24. JMLR.org, 2024.

[18] T. Miller, I. Durlik, E. Kostecka, P. Kozlovska, A. Łobodzińska, S. Sokołowska, and et al., "Integrating artificial intelligence agents with the internet of things for enhanced environmental monitoring: Applications in water quality and climate data," *Electronics*, vol. 14, 2025. [Online]. Available: https://www.mdpi.com/2079-9292/14/4/696

[19] V. Musanga, S. Viriri, and C. Chibaya, "A framework for integrating deep learning and symbolic ai towards an explainable hybrid model for the detection of covid-19 using computerized tomography scans," *Information*, vol. 16, 2025. [Online]. Available: https://www.mdpi.com/2078-2489/16/3/208

[20] C. Régis, J.-L. Denis, M. L. Axente, and A. Kishimoto, Eds., *Human-Centered AI: A Multidisciplinary Perspective for Policy-Makers, Auditors, and Users*. New York, NY: Routledge, 2025.