# Generative AI in E-maintenance: Myth or Reality?

J. Ćelić*, T. Bronzin**, M. Horvat***, A. Jović****, A. Stipić**, B. Prole**, M. Maričević*****, I. Pavlović*****, K. Pap*****, M. Mikota*****, N. Jelača*****

*University of Rijeka Faculty of Maritime Studies, Rijeka, Croatia
**CITUS, Zagreb, Croatia
***University of Zagreb Faculty of Electrical Engineering and Computing, Department of Applied Computing, Zagreb, Croatia
****University of Zagreb Faculty of Electrical Engineering and Computing, Dept. of. El., Microelectronics., Comp., Zagreb, Croatia
*****University of Zagreb/Faculty of Graphic Arts, Zagreb, Croatia

jasmin.celic@pfri.uniri.hr, tbronzin@citus.hr, marko.horvat3@fer.unizg.hr, alan.jovic@fer.unizg.hr, arian@citus.hr, brigita@citus.hr, marko.maricevic@grf.unizg.hr, ivana.pavlovic@grf.unizg.hr, klaudio.pap@grf.unizg.hr, miroslav.mikota@grf.unizg.hr, nina.jelaca@grf.unizg.hr

***Abstract*** — **With increasing requirements for reliability, availability, efficiency, effectiveness, productivity, and security of the system, the importance of diagnostics and maintenance is also increasing. E-maintenance as a leading concept for maintenance management has so far primarily involved the use of domain-specific technical language processing (TLP) techniques on historical case data. Due to its popularity, generative AI (GAI) with large language models (LLMs) is starting to be used more and more in various technical areas, thus starting to take an increasingly important place in diagnostics and maintenance. Starting from the fact that the rapid development of information and communication technologies (ICT) was the main factor in the emergence and development of the concept of e-maintenance, the importance of the potential more serious application of all forms of generative AI in the context is clear. This is especially pronounced in cases of difficult or impossible access to the location of components or an uncertain situation related to the type of process (e.g., nuclear, aeronautical, space, offshore). Autonomous vehicles, vessels, and aircraft (as an indispensable part of today's intelligent transport systems) are certainly a leading example of these cases. Regardless of the level of autonomy, these systems are extremely complex and difficult to maintain and represent a clear challenge for the application of new approaches. Therefore, the authors of the paper propose the use of middleware that would enable the integration of various GAI tools, algorithms, and models to increase the effectiveness of diagnostics and maintenance as close as possible to real-time. However, the exact extent of the possibilities and limitations of this approach has yet to be determined.**

***Keywords - diagnostics, maintenance, e-maintenance, AI, generative AI, intelligent transport systems, middleware***

## I. INTRODUCTION

In economy and industry, maintenance has always played an important role because it regularly represents a significant cost [1]. Effective maintenance concepts and strategies contributed to reducing such costs and mitigating business uncertainty. Malfunctions in the operation of devices and drives were primarily eliminated with the help of a corrective maintenance (CM) strategy, but this method did not affect the inevitability of work stoppages due to malfunctions and problems with the equipment. Therefore, it evolved into a much more proactive approach with the emergence of a preventive maintenance (PM) strategy that involved periodic inspections of devices and systems after a predefined period or a certain number of operating cycles. In this scenario, skilled technicians can proactively replace certain components prior to the occurrence of key faults or malfunctions in the system's operation that are likely to arise soon. Maintenance types based on EN 13306:2001 are shown on Fig. 1.

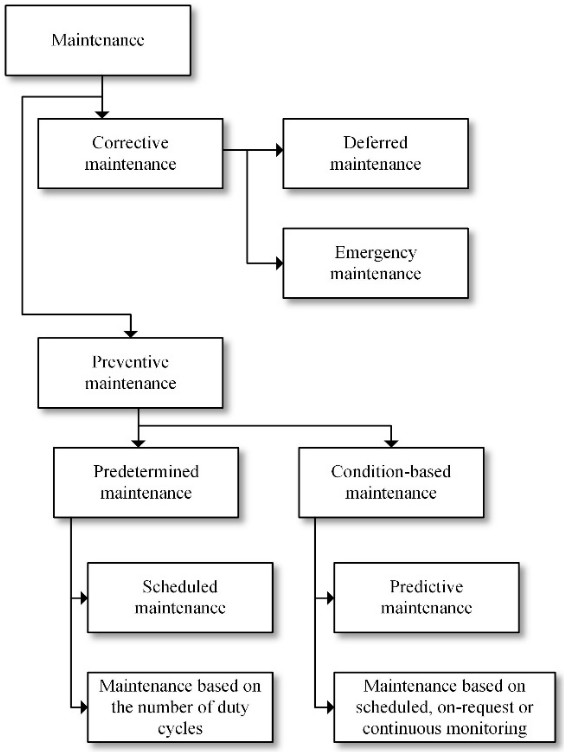


Figure 1. Maintenance types based on EN 13306:2001

With the development of primarily information and communication technologies (ICT) and sensors, a condition-based maintenance approach (CBM) appears [2]. Devices and systems are beginning to be monitored by sensors capable of measuring and monitoring certain physical quantities (such as temperature, pressure, force, sound, vibration, etc.) and processing the results independently in real time or forwarding them to other systems for analysis and decision-making. Depending on the measurement results and preset limit parameters, the staff in charge of maintenance will receive appropriate information and take appropriate measures based on it. Further technological development, especially in areas such as "big data", "data mining", "cloud computing", and "internet of things", brings more complex and effective approaches such as predictive maintenance (PdM) and remote maintenance (RM), respectively new strategies like e-maintenance. A PdM system is a cyber-physical system (CPS) in which networked assets can anticipate defects, identify their underlying causes, and autonomously detect problematic events [3]. Such a system can detect failures before they occur or predict the lifetime of a device component [4].

The need for a high level of automation and the transition to complete autonomy motivated engineers to create computer systems that enable remote maintenance and increasingly serious application of the e-maintenance strategy. Without such solutions, intelligent transport systems, such as partially or fully autonomous vehicles, vessels, or aircraft, would be impossible to maintain during their duties [5]. The lack of a crew or a small number of service and technical personnel is a serious problem in such systems. If there are even people in charge of maintenance (e.g. on semi-autonomous ships), their knowledge and skills are limited due to the diversity and complexity of the applied systems. A high-quality communication connection and a suitable computer system that will process the obtained information are required to quickly acquire professional and expert knowledge at a distance. Therefore, it is not unusual that soon after the appearance of the PdM system, various AI algorithms began to be used, further improving them [5], [7], [8].

Artificial intelligence algorithms can detect patterns associated with faults and failures or detect degradation early to implement appropriate countermeasures [4], [6]. Commonly used AI algorithms in PdM, Fig. 2, are from the machine learning (ML) and deep learning (DL) families.

ML provides machines with the ability to automatically learn, and act based on previous experience. The main types of machine learning algorithms are supervised, semi-supervised, unsupervised, and reinforcement-based learning. From the mentioned groups of algorithms for the PdM context, the following representatives can be singled out: artificial neural network (ANN) [7–9], decision tree (DT) [10], [11], support vector machine (SVM) [12], [13], k-nearest neighbors (k-NN) [14], [15], particle filter [16], principle component analysis [17], adaptive resonance theory [18] and self-organizing maps [19].

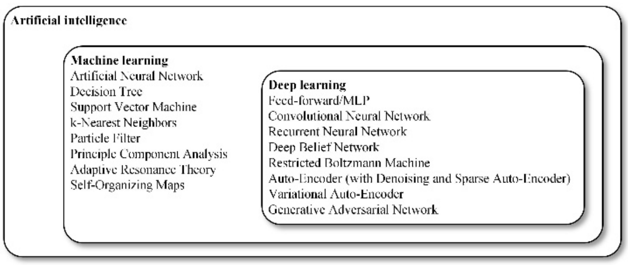


Figure 2. Commonly used AI algorithms in PdM

DL is a subset of ML that uses a neural network like the human neural system to mine data and analyze various factors. DL models outperform statistical and traditional ML models in many fields, including PdM, when sufficient historical data exists. DL has demonstrated its dominance in the field of PdM through its exceptional proficiency in feature learning, fault classification, and fault prediction using multilayer nonlinear transformations.

In the PdM field, DL models such as Feed-forward/MLP [20], convolutional neural network (CNN) [21], recurrent neural network (RNN) [22], [23], deep belief network (DBN) [24], restricted Boltzmann machine (RBM) [25], auto-encoder (AE) [26] with its adaptations denoising auto-encoder (DAE) [27] and sparse auto-encoder (SAE) [28], variational auto-encoder (VAE) [29] and generative adversarial network (GAN) [30] are most commonly used.

At the beginning of the 21st century, Generative AI (GAI) was developed from ML, representing a class of algorithms capable of generating new data. The previously mentioned GAN and large language models (LLM) represent an important subset of GAI. While GANs consist of two neural networks, a generator that produces synthetic data and a discriminator that distinguishes that data from real instances, LLMs generate human-like text by predicting the probability of a word given previous words used in the text. A special type of LLM is the globally known generative pre-trained transformer (GPT). GPT models are trained on massive amounts of textual data and can generate coherent, contextually relevant sentences.

Although LLM is key to GAI, it is important to mention that it is only a part of natural language processing (NLP), representing a broader field within artificial intelligence and computational linguistics. It focuses on enabling machines to understand, interpret, and create human language. NLP encompasses tasks such as sentiment analysis, machine translation, text mining, text summarization, named entity recognition, part-of-speech (POS) tagging, etc. Although NLP models trained on general language may not be directly applied to industrial text, promising progress was made in adapting NLP techniques to technical domains, which caused special interest in the industry. It was also reflected in the maintenance itself. This is primarily due to the technical language processing (TLP) engineering approach, which intelligently and selectively uses NLP tools for technical language data [31]. In this way, serious preconditions were created for using GAI within the framework of PdM and the e-maintenance concept.

## II. OPPORTUNITIES AND CHALLENGES OF E-MAINTENANCE

Over the past few decades, numerous organizational strategies have been established to optimize the efficiency of maintenance operations. Several approaches, such as total productive maintenance (TPM), reliability-centered maintenance (RCM), and condition-based maintenance, have been identified by researchers [32]. The previously mentioned approaches have been successfully adopted in the industry, yielding predominantly favorable outcomes.

Further technological development, especially in ICT, ensured the prerequisites for adopting new approaches and concepts in maintenance. The most effective approach involves maintaining constant awareness of the state of the property and its components and carrying out repair and servicing measures only when necessary. New wireless techniques and the Internet offer the possibility of using mobile devices to access large information globally and online [33], [34], [35].

Increasing demands for reliability, availability and security and the need to obtain asset information at any place and at any time cause the creation of new solutions in asset maintenance. One of these solutions is precisely the e-maintenance concept, which, among other things, provides a network that integrates and synchronizes various maintenance and reliability applications to collect and deliver asset information where it is needed [36].

The e-maintenance concept integrates telemetric maintenance principles with web services and modern e-collaboration methods. This form of collaboration enables sharing and exchanging information, knowledge, and e-intelligence [37], [38]. Starting from the fact that the rapid development of information and communication technologies was the main factor in the creation and development of the concept of e-maintenance, the importance of the potential more serious application of all forms of generative AI in the sense is recognizable. This is especially pronounced in cases of difficult or impossible access to the location of components or an uncertain situation related to the type of process (e.g., nuclear, aeronautical, space, offshore). Autonomous vehicles, vessels, and aircraft, an indispensable part of today's intelligent transport systems, are certainly a leading example of these cases [67]. Regardless of the level of autonomy, these systems are extremely complex and difficult to maintain and represent a clear challenge for the application of new approaches.

The integration of e-maintenance in modern industry presents both opportunities and challenges. The need for effective communication, collaboration, and change management strategies to ensure the alignment of data-driven solutions among stakeholders and users is emphasized [39]. These findings underscore the importance of a proficient workforce and appropriate tools and methodologies in successfully utilizing AI in e-maintenance.

E-maintenance can help reduce maintenance costs by identifying potential problems before they become major problems, enabling proactive maintenance, and reducing downtime. It can increase efficiency and streamline existing maintenance processes by using data analytics to predict when maintenance will be needed, reducing the time and resources required to perform maintenance tasks. This will lead to optimal equipment utilization and reduced downtime, increasing productivity and efficiency. Moreover, having real-time equipment performance data will contribute to making better decisions about maintenance schedules and resource allocation.

In addition to the wide range of opportunities e-maintenance provides, attention should be paid to the accompanying challenges. The effectiveness of e-maintenance relies heavily on the quality of data collected from the equipment's sensors. Poor data quality can lead to inaccurate predictions and unreliable decision-making. Interoperability caused by different standards and protocols and the need for integration with existing systems, infrastructure, and processes can create additional costs or completely limit functionality, so they need to be carefully planned. By changing business needs, e-maintenance systems must have the possibility of scalability without compromising reliability and stability. Very often, such systems for real-time analysis and high-quality and reliable processing of data sets also need powerful computer resources so that sophisticated algorithms can be executed quickly enough. Sometimes, very diverse technical and service personnel with different competencies, levels of knowledge, and skills must be ready to adapt to the upcoming new technologies and requirements, which certainly impacts their satisfaction level.

## III. Role of GAI in E-maintenance

Generative AI (GAI or G-AI) is a form of "limited memory AI" trained on a specific dataset and produces predictions by employing statistical analysis on that corpus of data. ChatGPT-3, for example, is trained on all publicly available data collections on the internet through September 2021. Generative AI algorithms are a class of AI algorithms that create new content in various forms (e.g., images, text, and music) by learning underlying patterns from training data. A much broader term than generative AI is AI-generated content (AIGC), which includes generative AI algorithms and other AI techniques such as natural language processing and computer vision. A large AI model refers to any neural network architecture that has many parameters, such as a large visual model (LVM), a large language model (LLM), and a large multimodal model [40]. AIGC refers to using generative AI algorithms to assist or replace humans in creating rich, personalized, and high-quality content at a faster pace and lower cost based on user inputs or requirements [41], [42].

Generative AI models are designed to acquire knowledge from extensive datasets, creating fresh content that closely resembles the original dataset. These models go beyond simple classification or forecasting jobs and create novel instances demonstrating artistic, intellectual, or other valuable attributes. Various categories of generative AI models exist, each specifically engineered to generate distinct forms of content:

- Variational Autoencoders (VAEs) are a class of generative models characterized by an encoder-decoder architecture [42], [43]. This approach employs a process of transforming input data into a latent space and subsequently reverts it to its original data domain. VAEs are characterized by the ability to achieve a trade-off between accurately reconstructing data and applying regularization techniques to generate new samples that conform to the learned data distribution.

  The training process of VAEs aims to optimize the model's parameters to reduce reconstruction errors and regularize the distribution of the latent space. The latent space representation enables the generation of unique and varied samples by changing the points inside it.

VAEs are utilized in diverse domains, including picture generation, anomaly detection, and data compression. They facilitate the creation of realistic visuals, artwork, and interactive exploration of latent regions. For instance, by training on real-time sensor readings, VAEs can capture complex data distributions and generate more robust predictions on future equipment conditions than traditional AI methods.

- Generative Adversarial Networks (GANs) operate through an adversarial process that consists of two primary elements: a generator network and a discriminator network. The primary goal of the generator is to provide data samples that are very persuasive, whereas the discriminator's role is to differentiate between authentic and artificially generated samples [44].

  The training process of GANs involves the generator attempting to confuse the discriminator while the discriminator aims to accurately classify the samples. During this dynamic competition, both networks enhance their performance over consecutive iterations.

  GANs have established a specialized position for themselves due to their exceptional contribution to picture synthesis. They enabled the creation of very realistic images, the alteration of artistic style, and the production of painted pictures. GANs have been utilized for generating images and transforming text into images, producing films, and generating realistic simulations for virtual environments.

  As the advancement and refinement of these generative models progress, one can anticipate the production of increasingly realistic and superior-quality AI-generated material. Nevertheless, using such models also gives rise to significant ethical and regulatory concerns, namely, copyright and data privacy matters.

- Auto-regressive models (ARMs) are notable in artificial intelligence for their distinctive methodology in generating new samples. These models estimate the conditional probability of each data point by considering the preceding context and relying on past values [44]. They operate sequentially, allowing for the formation of complex sequences.

  These models have been extensively utilized in text production, language modeling, and music composition. They possess exceptional proficiency in identifying interdependencies within sequences, producing coherent and contextually suitable outcomes.

- Flow-based models (FBMs) utilize the method of normalizing flows, which entails a series of reversible transformations, to accurately represent complex data distributions [44]. These changes facilitate efficient sampling and the computation of probabilities.

  They possess the capability to directly represent data distributions. They achieve this by establishing a reversible mapping between the input and output domains, enabling data production and efficient density estimation.

  Flow-based models are widely used in various domains, including picture production, density estimation, and anomaly detection. Notable strengths include the capacity to assess probability in a manageable manner, precise sampling, and the adaptability to represent the hidden space.

- Generative Diffusion Models (GDMs) are taught to restore clarity to images that have been blurred by Gaussian noise by effectively reversing the diffusion process [42], [45]. Multiple generative models based on diffusion have been suggested, such as diffusion probabilistic models, noise-conditioned scoring networks, and denoising diffusion probabilistic models.

- Transformer-based models (TBMs) are a distinct type of deep learning architecture that has gained significant popularity and effectiveness in Natural Language Processing (NLP).

  Transformer models are commonly used to construct language models, with one notable example being the Generative Pre-trained Transformer (GPT). The models in this line have demonstrated exceptional skill in producing contextually suitable and logically consistent language responses to a given instruction [46]. These models are extensively utilized in various NLP applications, including text completion, question answering, translation, and summarization.

Advanced behavioral analysis is applied by generative artificial intelligence to greatly enhance predictive maintenance in the manufacturing business [64], [65]. The procedure commences by conducting comprehensive data collection using many sensors put on the machines. These sensors continuously monitor several operational characteristics, including temperature, vibration, pressure, and acoustics. As a result, a substantial amount of data is gathered, providing valuable insights into the machine's operation and condition. The raw data is subjected to a feature extraction procedure, during which crucial patterns and characteristics are detected to provide significant insights into the performance and health of the system. A generative model, such as a Generative Adversarial Network (GAN) or Variation Autoencoder (VAE), is subsequently trained using these features. Training enables the model to comprehend the typical functioning patterns of the equipment, establishing a standard for identifying deviations from the norm.

After training, the model consistently analyzes real-time operational data and conducts behavioral analysis by comparing the real-time data to the taught 'normal' behavior. Any substantial variation suggests potential equipment malfunction, emphasizing the necessity for repair.

The system's capacity to identify prospective faults before equipment failure allows for implementing predictive maintenance strategies. Instead of following a conventional maintenance schedule, this technique schedules maintenance based on anticipated equipment failures, resulting in enhanced operational efficiency and optimal utilization of maintenance resources.

Furthermore, as the system gathers data over time, the generative model consistently acquires knowledge and improves its forecasts. Continuous learning enhances the precision of predictive maintenance capabilities.

Generative AI, with its behavior analysis and continuous learning, offers a novel and efficient method for predictive maintenance in production. This approach not only enhances productivity but also improves process transparency and product availability by greatly enhancing the overall efficiency and effectiveness of the production process.

Generative artificial intelligence has emerged as a revolutionary technology that has completely changed the existing scene in the field of artificial intelligence in recent years. Its ability to create and synthesize data has opened many possibilities in various domains. Condition monitoring and predictive maintenance (PdM) enabled the change of existing approaches and the improvement of existing equipment monitoring and maintenance practices.

In contrast to traditional artificial intelligence (AI) methodologies that rely on pre-established rules and labeled data to make predictions, Generative AI surpasses these approaches by generating novel data that adheres to similar patterns. By employing methodologies such as Generative Adversarial Networks (GANs), it comprehends and replicates the fundamental data patterns, facilitating the generation of innovative and previously unobserved data. This advanced technology has the potential to significantly transform maintenance methods, enabling a more proactive and efficient approach to equipment management. There exist numerous captivating uses of Generative AI, which encompass [66]:

- **Anomaly Detection**: The process of anomaly identification using generative models entails training these models to effectively reconstruct typical data patterns. Anomalies are identified during the testing phase by observing substantial variances in the reconstruction error. Such an approach exhibits superiority over conventional predictive anomaly detection techniques due to its unsupervised nature, eliminating the need for labeled data. Moreover, it demonstrates exceptional proficiency in identifying novel anomalies that deviate substantially from established normal patterns. Consequently, this approach showcases enhanced adaptability in real-world situations characterized by a scarcity of labeled anomaly instances.
- **Data augmentation**: involves the utilization of generative artificial intelligence models, such as Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs), to expand constrained datasets by generating additional samples that adhere to the fundamental patterns exhibited by the original data. This aspect holds significant value in scenarios where acquiring extensive labeled datasets poses difficulties. In the context of predictive maintenance, synthetic sensor data can be generated to mimic diverse equipment conditions and failure situations. This practice facilitates the thorough examination of algorithms and models through rigorous testing. Furthermore, the utilization of Generalized Artificial Intelligence (Gen AI) models has demonstrated its efficacy in effectively tackling the challenges associated with imbalanced data, particularly in scenarios of infrequent failures or anomalies. By synthesizing various instances of such scenarios, generative artificial intelligence (AI) effectively reduces bias within prediction models and improves their overall performance.
- **Uncertainty Estimation**: The issue of uncertainty estimates in predictive AI algorithms has been addressed by generative models. In contrast to conventional approaches that face challenges in accurately measuring uncertainty, generative artificial intelligence (AI) offers a viable solution by leveraging the latent space to generate several credible iterations of the data, hence facilitating reliable quantification of uncertainty. In critical domains such as predictive maintenance, this functionality boosts the dependability of decisions, enabling organizations to make more knowledgeable and assured choices by relying on accurately calibrated estimations of uncertainty.
- **Interpretability**: Certain generative models, such as Variational Autoencoders (VAEs), can generate a condensed representation of the input data, commonly known as the "latent space." This form of representation has the potential to offer valuable insights into the fundamental factors that contribute to anomalies, hence enhancing the interpretability of the results.
- **Multimodal-based Predictive Maintenance**: The integration of varied data sources, including sensor measurements, images, and textual information from manuals, guides, handbooks and troubleshooters, inspection notes, and other sources within the framework of generative AI, has the potential to significantly enhance the field of predictive maintenance. Complex, multimodal models provide a more extensive comprehension of equipment health, facilitating innovative forecasts, expedited anomaly identification, and enhanced precision in maintenance recommendations in contrast to conventional single-mode AI models.

Table I. shows an overview of generative AI models that can be used in PdM for different application areas.

TABLE I. APPLICATIONS OF GENERATIVE AI IN PREDICTIVE MAINTENANCE

| Applications | GAI models | References |
|---|---|---|
| Anomaly detection | VAE, GAN RF*, ML, SVM, NN | [30], [42], [43], [68], [69], [70] |
| Data augmentation | ML, DL, GAN, MLP, SVM, KNN | [71], [72], [73], [74], [12], [13], [14], [15], |
| Uncertainty estimation | CNN, LSTM*, QR*, KDE*, BNN | [21], [22], [23], [75], [76] |
| Interpretability | VAE, ML, DT, GB* | [10], [11], [77], [78] |
| Multimodal-based predictive maintenance | SVM, MKL*, DBN, RBM, DAE, CNN, RNN, GAN, MTL, MMTL | [79], [24], [25], [27], |

* RF - random forest; LSTM - long short-term memory; QR - quantile regression; KDE - kernel density estimation; GB – gradient boosting; MKL – multi kernel learning; MTL – multitask learning; MMTL – multimodal transfer learning

The introduction of generative artificial intelligence (AI) in preventive maintenance, particularly in condition monitoring and predictive maintenance, offers promising prospects. However, this integration also brings up distinct dangers and challenges. The intricate nature of generative artificial intelligence (AI) algorithms requires substantial computational resources, underscoring the importance of strategic planning and investments in infrastructure to ensure the effective implementation of these algorithms. Furthermore, ensuring the protection of data privacy and security assumes great importance, especially in the context of managing confidential equipment information and maintenance logs. In addition, it is imperative to consider and handle ethical and privacy considerations about sensitive data, as this plays a vital role in upholding trust and ensuring adherence to regulations during the entirety of the implementation procedure.

## IV. Needs for Middleware

Middleware facilitates smooth communication and interaction among various software components, operating systems, and devices. It offers a uniform platform that conceals the intricacy of the underlying systems and facilitates effective application development and integration. Middleware must tackle precise engineering obstacles, such as creating mobile applications and incorporating outdated technologies into novel services [47]. The system should possess several essential features, including the capacity to manage remote applications, compatibility with various communication protocols, effective memory management, and a high level of abstraction for application designers. Furthermore, the middleware must be able to expand and adjust to various settings, such as ad-hoc network setups. In addition, middleware must offer security protocols to safeguard confidential information shared between applications, guarantee dependable communication and resilience to errors, and facilitate compatibility across many platforms and technologies. Middleware requirements encompass the need for smooth communication and interaction, standardization, resolution of engineering challenges, support for distributed applications and diverse communication protocols, efficient memory management, a high level of abstraction, scalability and adaptability to different environments, security measures, reliable communication, and fault tolerance, as well as interoperability with various platforms and technologies.

Middleware plays a vital role in contemporary software architecture by enabling communication between diverse applications and systems. Given the growing utilization of AI in maintenance, middleware can have a crucial function in seamlessly incorporating AI models into current systems.

Data processing and filtering are crucial functions of middleware in maintenance aided by AI. Middleware facilitates data preprocessing from diverse sources, including sensors and logs, to guarantee that only pertinent and refined data is supplied to AI models. Additionally, it can perform data transformation, which involves changing data from one format to another to fulfill the specifications of various AI models.

Middleware cannot only process data but also carry out authentication and permission functions to guarantee that only authorized individuals may access AI models and their outputs. To optimize the utilization of AI models in maintenance, the middleware must possess the capability to efficiently process substantial amounts of data and handle multiple requests concurrently without experiencing substantial delays.

Scalability is another important requirement for middleware in this context. It should be able to effectively manage growing volumes of data traffic as maintenance activities expand. To seamlessly interact with preexisting systems, middleware must possess standardized Application Programming Interfaces (APIs) and be compatible with various operating systems. Middleware solutions must provide the flexibility to adapt to evolving requirements in maintenance operations and scale proportionally.

Middleware in maintenance backed by AI models must possess robust error-handling capabilities as a crucial requirement. Middleware should be developed to effectively handle such circumstances and offer informative error messages for technical support teams. To guarantee transparency and traceability of AI model's predictions and actions, the middleware must record all transactions and offer comprehensive analytics. To adhere to security standards and regulations, middleware must possess strong security attributes, including encryption, authentication systems, and input validation.

Furthermore, as the system gathers data over time, the generative model consistently acquires knowledge and improves its forecasts. Continuous learning enhances the precision of predictive maintenance capabilities.

Generative AI, with its behavior analysis and continuous learning, offers a novel and efficient method for predictive maintenance in production. This approach not only enhances productivity but also improves process transparency and product availability by greatly enhancing the overall efficiency and effectiveness of the production process.

Middleware is crucial for optimizing data processing and facilitating the management of AI models in the context of predictive maintenance. Optimal utilization of AI models necessitates data processing in a manner that considers the specific context in which it is being used [48]. Collecting and managing data in real-time is a highly significant work in complex systems, and it can be made easier using middleware [49]. Additional authors [50], [51] also highlight the necessity of middleware to effectively link components and facilitate the development of intricate systems, a particularly crucial aspect in predictive maintenance. Collectively, these studies emphasize the crucial function of middleware in enhancing the efficiency of utilizing AI models in maintenance operations.

## V. Future Directions

The future of AI in predictive maintenance is promising, with a focus on renewable energy systems [52], industrial systems monitoring [53], and intelligent maintenance frameworks [54]. Machine learning and deep learning algorithms are being explored for failure detection and classification in high-performance computing systems [55].

The shift from traditional maintenance to predictive maintenance is driven by smart manufacturing and IoT [56]. Quality management systems are being redefined in the Industry 4.0 era, leveraging AI and big data analytics for predictive quality management [57]. The application of AI in predictive maintenance is exemplified by the successful prediction of machine failure using an artificial neural network [58]. The use of deep learning models in predictive maintenance is also being explored, focusing on anomaly detection, root cause analysis, and remaining useful life estimation [59].

Current generative AI models are often considered black boxes because they lack transparency and interpretability, making it a challenge to understand their decision-making processes [40]. These concerns have raised worries about outcomes, security threats, and ethical problems. As a result, there is a growing need to create explanatory models for GAI. Designing explainable GAI models involves the utilization of various important technologies, such as interpretable artificial intelligence algorithms, model visualization tools [60], and human-in-the-loop (HITL) approaches [61]. Machine learning algorithms that are interpretable, such as decision trees and rule-based models, allow for capturing intricate correlations between input features and outputs. As a result, these algorithms explain the model's output. Furthermore, model visualization approaches [60], such as activation mapping, saliency maps, and feature visualization, provide a visual representation of the decision-making process of GAI models. This helps users comprehend how models classify, group, or connect various inputs. HITL approaches incorporate the participation of human specialists in the decision-making process. GAI models can enhance this by jointly designing interfaces and implementing interactive feedback mechanisms to get improved outcomes [61]. By integrating these technologies, the transparency and interpretability of GAI models can be enhanced [40].

One of the promising technologies is that a digital twin usually consists of a simulation model that will be continuously updated to reflect the state of their real-life twin. These emerging frameworks enable the acquisition of extensive data regarding the failure performance of crucial and pertinent components. Effective execution of error detection and prediction is highly advantageous and essential. A specific instance is a way of diagnosing faults in two phases, which digital twins facilitate through deep transfer learning [62].

Moreover, technologies like Virtual Reality (VR) and Augmented Reality (AR) can replicate essential procedures and conduct virtual assessments of manufacturing lines and machinery. This aids in identifying errors that may result in prospective interruptions and resolving them before impeding operations. VR/AR can be utilized to instruct maintenance technicians in a virtual setting and enable them to rehearse and acquire new proficiencies without the necessity of tangible apparatus or actual hazards. In addition, technicians can obtain immediate remote support from professionals situated in different locations, enhancing efficiency and minimizing the necessity for physical visits, which is particularly important in the ITS domain. Moreover, this technology can be employed to carry out virtual examinations of equipment and machinery, enabling technicians to detect prospective faults before they escalate into significant complications. VR/AR technology-enabled interactive manuals facilitate maintenance procedures by offering technicians clear and sequential instructions, simplifying the comprehension and execution of intricate jobs. VR/AR technology enables interactive manuals that offer technicians step-by-step instructions for maintenance operations, facilitating comprehension and completion of intricate jobs. It facilitates remote collaboration among technicians, allowing them to share information and work together more efficiently, enhancing teamwork and minimizing errors. Virtual reality (VR) and augmented reality (AR) technologies can decrease expenses related to maintenance training, inspections, and repairs by minimizing the requirement for physical prototypes, travel, and on-site visits.

All mentioned future directions could be combined with appropriate middleware that simplifies the process of connecting applications that were not designed to connect and provides functionality to connect them intelligently, streamlining application development and speeding time to market. AI middleware is software that enables communication and connection between applications or application components in a distributed network [63]. All this implies its use in the field of GAI, whose potential should further improve the concept of e-maintenance, especially in the field of poorly accessible or inaccessible complex systems.

## VI. Conclusions

The ability of intelligent transport systems to act adaptively in changing conditions and situations based on data collected and processed in real-time enables information transparency, controllability, and improved response of transport systems, which ultimately makes them intelligent. The complexity of such systems is usually very high, so it is understandable why they require special maintenance approaches. This is especially evident in the maintenance of dislocated and hard-to-reach systems. E-maintenance and PdM have proven to be satisfactory solutions in such cases.

The transition from the now traditional maintenance methods to the dynamic realm of GAI represents a significant leap in efficiency and effectiveness. GAI will only be better and more advanced in future versions, and with increasingly extensive TLP resources and the use of appropriate language models, the success of this approach will be guaranteed. Therefore, the future of maintenance lies in the adaptive learning capabilities of platforms and middleware powered by GAI. Although the use of GAI in e-maintenance is becoming more certain every day, it must not be forgotten that AI decision-making comes with AI responsibility and that AI systems must also be maintained.

## Acknowledgment

This work was supported under the project “AIM - Research and the initial development of middleware connectivity with AI and GAI services”, based on the program “Targeted scientific research”, reference number

C3.2.R3-I1.04 issued by the Republic of Croatia, Ministry of Science and Education.

## References


[1] M. Bevilacqua, M. Braglia, “Analytic Hierarchy Process Applied to Maintenance Strategy Selection. Reliability Engineering and System Safety”, 70/ 1: 71–83, 2000

[2] J. Lee, H.A. Kao, S. Yang, “Service innovation and smart analytics for Industry 4.0 and big data environment”. Vol. 16 Procedia CIRP Elsevier: 3–8, 2014

[3] J. Lee, C. Jin, B. Bagheri, “Cyber physical systems for predictive production systems”, Prod. Eng., 11, 155–165., 2017

[4] G. B. Huang, H. Lee, E. Learned-Miller, “Learning hierarchical representations for face verification with convolutional deep belief networks”, In 2012 IEEE conference on computer vision and pattern recognition, IEEE, pp. 2518–2525, 2012

[5] A.Theissler, J. Pérez-Velázquez, M.Kettelgerdes, G. Elger, “Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry”, Reliability Engineering & System Safety, Volume 215, 2021

[6] S. Sajid, A. Haleem, S. Bahl, M. Javaid, T. Goyal, M. Mittal, “Data science applications for predictive maintenance and materials science in context to industry 4.0.”, Mater Today Proc 2021

[7] A. Sharma, R. Jigyasu, L. Mathew, S. Chatterji, “Bearing fault diagnosis using frequency domain features and artificial neural networks. Information and Communication Technology for Intelligent Systems”, Springer, pp. 539–547, 2019

[8] M. Elforjani, S. Shanbr, “Prognosis of bearing acoustic emission signals using supervised machine learning”, IEEE Transactions on industrial electronics, vol. 65, no. 7, pp. 5864–5871, 2017

[9] I. M. Karmacharya, R. Gokaraju, “Fault location in ungrounded photovoltaic system using wavelets and ann”, IEEE Transactions on Power Delivery, vol. 33, no. 2, pp. 549–559, 2017

[10] L. Kou, Y.Qin, X. Zhao, Y. Fu, “Integrating synthetic minority oversampling and gradient boosting decision tree for bogie fault diagnosis in rail vehicles”, Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, vol. 233, no. 3, pp. 312–325, 2019

[11] S. S. Patil and V. M. Phalle, “Fault detection of anti-friction bearing using adaboost decision tree,” in Computational Intelligence: Theories, Applications and Future Directions-Volume I. Springer, pp. 565–575, 2019

[12] X. Zhu, J. Xiong, Q. Liang, “Fault diagnosis of rotation machinery based on support vector machine optimized by quantum genetic algorithm”, IEEE Access, vol. 6, pp. 33 583–33 588, 2018

[13] H. Han, X. Cui, Y. Fan, H. Qing, “Least squares support vector machine (ls-svm)-based chiller fault diagnosis using fault indicative features,” Applied Thermal Engineering, 2019.

[14] D. K. Appana, M. R. Islam, J.-M. Kim, “Reliable fault diagnosis of bearings using distance and density similarity on an enhanced knn,” in Australasian Conference on Artificial Life and Computational Intelligence, pp. 193–203. Springer, 2017

[15] S. R. Madeti, S. Singh, “Modeling of pv system based on experimental data for fault detection using knn method”, Solar Energy, vol. 173, pp. 139–151, 2018

[16] N. Daroogheh, N. Meskin, K. Khorasani, “A dual particle filterbased fault diagnosis scheme for nonlinear systems”, IEEE transactions on control systems technology, vol. 26, no. 4, pp. 1317–1334, 2018

[17] X. Deng, X. Tian, S. Chen, C. J. Harris, “Nonlinear process fault diagnosis based on serial principal component analysis”, IEEE transactions on neural networks and learning systems, vol. 29, no. 3, pp. 560–572, 2018

[18] X. Liu, H. Liu, J. Yang, G. Litak, G. Cheng, S. Han, “Improving the bearing fault diagnosis efficiency by the adaptive stochastic resonance in a new nonlinear system”, Mechanical Systems and Signal Processing, vol. 96, pp. 58–76, 2017

[19] A. Rai, S. H. Upadhyay, “Intelligent bearing performance degradation assessment and remaining useful life prediction based on selforganising map and support vector regression”, Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, vol. 232, no. 6, pp. 1118–1132, 2018

[20] P. J. Werbos, “Applications of advances in nonlinear sensitivity analysis. In System Modeling and Optimization”, Springer, 762–770, 2005

[21] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel.“Backpropagation Applied to Handwritten Zip Code Recognition”, Neural Computation 1, 4, pp. 541–551, 1989

[22] K. Cho, B. van Merriënboer, D. Bahdanau, Y. Bengio, “On the Properties of Neural Machine Translation: Encoder–Decoder Approaches”, In Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. Association for Computational Linguistics, Doha, Qatar, pp.103–111, 2014

[23] S. Hochreiter, J. Schmidhuber, “Long short-term memory”, Neural computation 9, 8, pp. 1735–1780, 1997

[24] G. E. Hinton, R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks”, Science 313, 5786, pp. 504–507, 2006

[25] C. Scholar, P. Smolensky, “Parallel distributed processing: Explorations in the microstructure of cognition. Information”, Processing in Dynamical Systems: Foundations of Harmony Theory, MIT Press, Cambridge, MA, USA 1, 667, pp. 194–281, 1986

[26] D. H. Ballard, “Modular Learning in Neural Networks” Aaai pp. 279–284, 1987

[27] Y. L. Cun, F. Fogelman-Soulié, “Modèles connexionnistes de l’apprentissage”, Intellectica. Revue de l’Association pour la Recherche Cognitive 2, 1, pp. 114–143, 1987

[28] A. Makhzani, B. Frey, “K-sparse autoencoders”, arXiv preprint, arXiv:1312.5663, 2013

[29] D. P. Kingma, M. Welling, “Auto-encoding variational bayes”, In 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings, Vol. 1. arXiv:1312.6114, 2014

[30] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio,” Generative adversarial nets”, In Advances in Neural Information Processing Systems, Vol. 3., pp. 2672–2680, 2014

[31] M. P. Brundage, T. Sexton, M. Hodkiewicz, A. Dima, S. Lukens, “Technical language processing: Unlocking maintenance knowledge”, Manufacturing Letters, 27, pp. 42–46, 2021

[32] A.C. Márquez, “The maintenance management framework”, Springer, London, 2007

[33] K. Holmberg, A. Helle, “Tribology as basis for machinery condition diagnostics and prognostics”, International Journal of Performability Engineering 4, pp. 255–269, 2008

[34] B. Iung, E. Levrat, A.C. Marquez, H. Erbe, “E-maintenance: principles, review and conceptual framework”, IFAC Proceedings Volumes, Volume 40, Issue 19, pp. 18-29, 2007

[35] E. Levrat, B. Iung, A. C. Marquez, “E-maintenance: review and conceptual framework”, Production Planning & Control, 19:4, pp. 408-429, 2008

[36] R. C. Baldwin, “Enabling an e-Maintenance infrastructure”, Maintenance Technology 12, 2001

[37] A. Muller, A.C. Márquez, B.Iung, “On the concept of e-maintenance: Review and current research”, Reliability Engineering and System Safety 93, pp. 1165–1187, 2008

[38] B. Iung, A.C. Márquez, “A review of e-maintenance capabilities and challenges”, 2008

[39] C. Emmanouilidis, E. Jantunen, E. Gilabert, A. Arnaiz, A. Starr, “e-Maintenance Update: The Road to Success for Modern Industry”, Proceedings of the 24th International Conference on Condition Monitoring and Diagnostic Engineering Management, COMADEM 2011, Stavanger, Norway, 2011

[40] Y.Wang, Y. Pan, M. Yan, Z. Su, T.H. Luan, “A Survey on ChatGPT: AI-Generated Contents, Challenges, and Solutions”, IEEE Open Journal of the Computer Society. pp. 1-20, 2023

[41] J. Wu, W. Gan, Z. Chen, S. Wan, H. Lin, “AI-generated content (AIGC): A survey,” arXiv preprint arXiv:2304.06632, pp. 1–17, 2023

[42] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, L. Sun, “A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT,” arXiv preprint arXiv:2303.04226, pp. 1–44, 2023

[43] G. Harshvardhan, M. K. Gourisaria, M. Pandey, S. S. Rautaray, “A comprehensive survey and analysis of generative models in machine learning,” Comput. Sci. Rev., vol. 38, p. 100285, 2020

[44] T. Karras, T. Aila, S. Laine, J. Lehtinen, “Progressive growing of gans for improved quality, stability, and variation,” in Proc. Int. Conf. Mach. Learn., 2018

[45] P. Dhariwal, A. Nichol, “Diffusion models beat GANs on image synthesis,” Adv. Neural Inf. Process. Syst., vol. 34, pp. 8780–8794, 2021.

[46] A. Gillioz, J. Casas, E. Mugellini, O. A. Khaled, "Overview of the Transformer-based Models for NLP Tasks," 15th Conference on Computer Science and Information Systems (FedCSIS), Sofia, Bulgaria, 2020, pp. 179-183, 2020

[47] E. Hoos, “Design method for developing a mobile engineering-application middleware (meam)”, 2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS), 2014

[48] M. P. Jensen, “IoT-middleware requirements for context-sensitive processing of data to enable predictive maintenance through augmented reality”, 2021

[49] T. Coito, “A Middleware Platform for Intelligent Automation: An Industrial Prototype Implementation.” Comput. Ind. 123: 103329, 2020

[50] J.W. Chang, K.Yong-Ki. “Design and Implementation of Middleware and Context Server for Context-Awareness”, IEEE International Conference on High Performance Computing and Communications, 2006

[51] Y. Georgalis, “Middleware for Ambient Intelligence Environments: Reviewing Requirements and Communication Technologies.” Interacción, 2009

[52] Y. S. Afridi, “Artificial intelligence based prognostic maintenance of renewable energy systems: A review of techniques, challenges, and future research directions.” International Journal of Energy Research 46, pp. 21619 – 21642, 2021

[53] A. Esteban, “Data mining in predictive maintenance systems: A taxonomy and systematic review”, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 12, 2022

[54] H. Zheng, “Advancing from Predictive Maintenance to Intelligent Maintenance with AI and IIoT”, ArXiv abs/2009.00351, 2020

[55] A. L. C. D. Lima,, “Smart predictive maintenance for high-performance computing systems: a literature review”, The Journal of Supercomputing 77, pp. 13494 – 13513, 2021

[56] Y. Ran, “A Survey of Predictive Maintenance: Systems, Purposes and Approaches”, ArXiv abs/1912.07383, 2019

[57] S.M. Lee, “The quality management ecosystem for predictive maintenance in the Industry 4.0 era”, International Journal of Quality Innovation 5, pp. 1-11, 2019

[58] H. Kamel, “Artificial intelligence for predictive maintenance.” Journal of Physics: Conference Series 2299, 2022

[59] O. Serradilla, “Deep learning models for predictive maintenance: a survey, comparison, challenges and prospect.” ArXiv abs/2010.03207, 2020

[60] A. B. Arrieta, N. D´ıaz-Rodr´ıguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garc´ıa, S. Gil-L´opez, D. Molina, R. Benjamins ,“Explainable artificial intelligence XAI: Concepts, taxonomies, opportunities and challenges toward responsible AI,” Information fusion, vol. 58, pp. 82–115, 2020

[61] T. N. Nguyen, R. Choo, “Human-in-the-loop XAI-enabled vulnerability detection, investigation, and mitigation,” in Proceedings of IEEE/ACM International Conference on Automated Software Engineering (ASE), pp. 1210–1212, 2021

[62] Y. Xu, Y. Sun, X. Liu, Y. Zheng, “A digital-twin-assisted fault diagnosis using deep transfer learning,” IEEE Access, vol. 7, pp. 19 990–19 999, 2019

[63] A. Farahzadi, P. Shams, J. Rezazadeh, R. Farahbakhsh, "Middleware technologies for cloud of things: a survey", Digital Communications and Networks, Volume 4, Issue 3, pp.176-188, 2018

[64] E. T. Bekar, P. Nyqvist and A. Skoogh, “An intelligent approach for data pre-processing and analysis in predictive maintenance with an industrial case study”, Advances in Mechanical Engineering 12, n. pag., 2020

[65] M. Nacchia, F. Fruggiero, A. Lambiase and K. Bruton, “A Systematic Mapping of the Advancing Use of Machine Learning Techniques for Predictive Maintenance in the Manufacturing Sector”, Applied Sciences, n. pag., 2021

[66] R. Gozalo-Brizuela, E. C. Garrido-Merch'an, “A survey of Generative AI Applications”, ArXiv abs/2306.02781, n. pag., 2023

[67] L. Stappen, J. Dillmann, S. S. Striegel, H. J. Vögel, N. Flores-Herr and B. W. Schuller, “Integrating Generative Artificial Intelligence in Intelligent Vehicle Systems”, ArXiv abs/2305.17137, n. pag., 2023

[68] J. Sharma, M. L. Mittal, G. Soni, "Explainable artificial intelligence (XAI) enabled anomaly detection and fault classification of an industrial asset", Recent Patents on Engineering, Volume 18, Issue 5, 2024

[69] S. R. Bhujbal, "Predictive Machine Maintenance Using Tiny ML", International Journal For Science Technology And Engineering, n. pag., 2023

[70] U. Albertin, G. Pedone, M. Brossa, G. Squillero, M. Chiaberge, "A Real-Time Novelty Recognition Framework Based on Machine Learning for Fault Detection", Algorithms, n. pag., 2023

[71] D. Dhanraj, A. Sharma, G. Kaur, S. Mishra, P. Naik and A. Singh, "Comparison of Different Machine Learning Algorithms for Predictive Maintenance", International Conference for Advancement in Technology (ICONAT), pp. 1-7, 2023

[72] S. M. Hallaji, Y. Fang, B. K. Winfrey, "Data augmentation for deep learning-based detection of pump anomalous conditions", IOP Conference Series: Earth and Environmental Science, n. pag., 2022

[73] V. Bui, T. L. Pham, H. Nguyen, Y. M. Jang, "Data Augmentation Using Generative Adversarial Network for Automatic Machine Fault Detection Based on Vibration Signals", Applied Sciences 11 (5), n. pag., 2021

[74] N. Pius, N. Innocent, "Big Data Analytics for Predictive System Maintenance Using Machine Learning Models", Advances in data science and adaptive analysis (vol 15), n. pag., 2023

[75] C. Chuang, T. Guanye, S. Jian-gang, S. Mouquan, Z. H. Zheng, "A Lithium-Ion Battery Degradation Prediction Model with Uncertainty Quantification for Its Predictive Maintenance", IEEE transactions on industrial electronics, n. pag., 2023

[76] L. Zhuang, A. Xu, X. L. Wang, "A prognostic driven predictive maintenance framework based on Bayesian deep learning", Reliability Engineering & System Safety, Vol 234, n. pag., 2023

[77] M. Amram, J. Dunn, J. J. Toledano, Y. D. Zhuo, "Interpretable predictive maintenance for hard drives", Machine Learning with Applications, Vol 5, n. pag., 2021

[78] B. Steurtewagen, D. V. Poel, "Adding interpretability to predictive maintenance by machine learning on sensor data", Computers & Chemical Engineering, Vol 152, n. pag., 2021

[79] K. Bayoudh, R. Knani, F. Hamdaoui, et al. "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets", Visual Computer, vol 38, 2939–2970, 2022